Econometric Theory, **0**, 2025, 1–39.
doi:10.1017/S0266466624000343

LEAST TRIMMED SQUARES: NUISANCE PARAMETER FREE ASYMPTOTICS

Vanessa Berenguer-Rico
University of Oxford

Bent Nielsen
University of Oxford

The Least Trimmed Squares (LTS) regression estimator is known to be very robust to the presence of "outliers". It is based on a clear and intuitive idea: in a sample of size n, it searches for the h-subsample of observations with the smallest sum of squared residuals. The remaining $n-h$ observations are declared "outliers". Fast algorithms for its computation exist. Nevertheless, the existing asymptotic theory for LTS, based on the traditional ϵ-contamination model, shows that the asymptotic behavior of both regression and scale estimators depend on nuisance parameters. Using a recently proposed new model, in which the LTS estimator is maximum likelihood, we show that the asymptotic behavior of both the LTS regression and scale estimators are free of nuisance parameters. Thus, with the new model as a benchmark, standard inference procedures apply while allowing a broad range of contamination.

1. INTRODUCTION

The Least Trimmed Squares (LTS) estimator (Rousseeuw, 1984) is known to be very robust to outliers. The robustness of the LTS estimator is often expressed through its high breakdown point (Rousseeuw and Leroy, 1987, Sect. 3.4). This means that the estimator remains bounded if, for a given sample, we distort nearly half of the observations in an arbitrary way. This contrasts with the OLS estimator and the quantile regression estimator, which have breakdown point close to zero (He et al., 1990), so that adding one observation to a sample may change those estimators unboundedly. Another attractive feature is that LTS is scale equivariant just as OLS and quantile regression, but in contrast to other M-estimators. Our concern is how to conduct nuisance parameter free inference with LTS.

The LTS estimator is computed as follows. The user specifies that a sample with n observations has h "good" observations and $n-h$ "outliers". The LTS

Comments from P. Guggenberger and the referees are gratefully acknowledged. Address correspondence to Vanessa Berenguer-Rico, Mansfield College & Department of Economics, University of Oxford, Oxford, UK, e-mail: vanessa.berenguer-rico@economics.ox.ac.uk

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

estimator is the OLS estimator for the h sub-sample with the smallest residual sum of squares. Thus, LTS gives an estimator of the unknown regression parameter and finds two different groups of observations: the “good” and the “outliers”. In a location-scale model this search is of linear order, while in regression it is of binomial order, hence, making analysis harder in the regression context.

The traditional approach in robust statistics is to analyze the asymptotic properties of the LTS estimator under the assumption that the regression errors are independent draws from a common ϵ-contaminated normal distribution, which mixes a normal distribution with a contamination distribution as popularized by Huber (1964). The errors are also assumed to be independent of the regressors and the parameter space is assumed to be compact. In this setting, asymptotic inference involves nuisance parameters depending on the contamination distribution (Rousseeuw, 1985; Croux and Rousseeuw, 1992). Butler (1982) gave a formal proof for the location-scale case. Čížek (2005) and Víšek (2006) considered the regression case with symmetric density and compact parameter space. A compact parameter space assumption often appears innocuous, but seems less appropriate when the purpose of robust estimation is to guard against arbitrary distortions in the presence of “outliers”. The common error distribution assumed in this stream of the literature is generally unknown and the practice is to apply nuisance parameters as if all observations are normal. For later reference, we term this approach as standard LTS, or in short SLTS.

Our analysis departs from this traditional approach in robust statistics and delivers an inferential theory that is free of nuisance parameters. Besides, it does not require a compact parameter space. Our asymptotic theory is inspired by the LTS model of Berenguer-Rico, Johansen, and Nielsen (2023), in which h “good” observations follow a classical regression model, while $n-h$ “outliers” have regression errors that are more extreme than the realized “good” errors. The LTS estimators for regression, scale and “outlier” classification maximize the ϵ-likelihood of the LTS model, in the sense of Scholz (1980). This was proved along with an asymptotic theory for the location-scale case. Although informative, the location-scale model is of limited applicability. Asymptotic theory for the regression case is the relevant theory for practitioners. We deliver this theory in this paper. The asymptotic arguments used by Berenguer-Rico et al. (2023) in the linear-order location-scale case do not immediately generalize to the regression case due to the complexity of the binomial search, which makes the theoretical problem much harder. We argue differently and more generally here delivering a nuisance parameter free asymptotic theory for the LTS regression estimator.

Specifically, the asymptotic analysis in the present paper starts by showing that the LTS estimator is bounded in probability. To the best of our knowledge, this result is new in the literature. Boundedness is derived under mild assumptions. The proof adapts a recent argument for M-estimators with non-convex criterion functions (Johansen and Nielsen, 2019).The boundedness result resonates with

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

the high breakdown point property of the LTS estimator and avoids a compact parameter space assumption. Next, we show that the proportion of "good" observations is consistently selected and derive the rate at which this consistent selection occurs. This uses theory on extreme and intermediate quantiles (Chibisov, 1964; Galambos, 1978; Leadbetter, Lindgren, and Rootzén, 1982). The final result is an asymptotic expansion of the LTS estimator in terms of the infeasible OLS estimator for the "good" observations. This asymptotic equivalence between the LTS estimator and the infeasible OLS estimator is shown for both the regression parameters and the scale estimators. This result shows that, in contrast to the traditional approach based on the ϵ-contamination model, no nuisance correction and consistency factors are required. The usual asymptotic distribution theory for OLS estimators then applies under different assumptions to the "good" observations such as i.i.d. or heteroscedastic structures and stationary or non-stationary time series regressors.

In simulations, we consider OLS, LTS, and SLTS inference for various contaminated samples. The simulations confirm the asymptotic theory and the fact that the underlying model is of primary importance when conducting inferences with the LTS estimator.

In practice, the user will have to choose the number h of "good" observations. A related matter is to decide whether LTS or SLTS inference is applicable. An estimator for h and model choice are discussed in Berenguer-Rico et al. (2023). An overview of these practical aspects is given in Section 6 below.

Many extensions of LTS exist in the literature: covariance estimation using the minimum covariance determinant (MCD) estimator (Rousseeuw, 1985), the Least Trimmed sum of Absolute deviations (LTA) estimator, which is a robust version of the Least Absolute Deviations (LAD) estimator (Hössjer, 1994), nonlinear regression in time series (Čížek, 2005), multivariate regression (Agullo, Croux, and Van Aelst, 2008), the forward search algorithm (Atkinson, Riani, and Cerioli, 2010), sparse regression (Alfons, Croux, and Gelper, 2013), canonical correlation analysis (Wilms and Croux, 2016), algorithms for fraud detection (Rousseeuw et al., 2019) or covariance estimation using cell-wise outliers (Raymaekers and Rousseeuw, 2023). Some of these estimators are examples of trimmed likelihood estimators (Bednarski and Clarke, 1993; Vandev and Neykov, 1993; Gallegos and Ritter, 2009; Clarke, 2018). The analysis presented here will be useful for analyzing and applying these extensions. More generally, the asymptotic analysis in this paper opens up for new inferential procedures in the presence of "outliers".

The paper is organized as follows. Section 2 describes the LTS estimator and the LTS model. Section 3 contains the asymptotic results: boundedness, consistent selection, and asymptotic expansion. Section 4 discusses regressors allowed by the theory. Section 5 illustrates the theory via simulations. Section 6 concludes with a discussion of some practical aspects of LTS. Proofs and technical derivations can be found in the Appendices.

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

2. THE LTS ESTIMATOR AND THE LTS MODEL

2.1. The LTS Estimator

We consider the linear regression for a scalar y_i and a vector x_{in} of regressors given by

$$y_i = x_{in}'\beta + \sigma\varepsilon_i \qquad \text{for } i = 1,\dots,n, \tag{2.1}$$

where x_{in} would usually include an intercept, but it does not have to. With this formulation of the model equation (2.1), all normalizations are built into the regressors x_{in} so that estimators for β will be $n^{1/2}$ consistent. For example, x_{in} could be an i.i.d. regressor, a level shift after a fraction of the sample $0 < \tau < 1$ so that $x_{in} = 1_{(i \le \tau n)}$, or a normalized random walk, $x_{in} = n^{-1/2}\sum_{\ell=1}^{i}\psi_\ell$ with i.i.d. increments ψ_ℓ. In the notation for y_i, we suppress the dependence on n noting that in the asymptotic analysis y_i is always replaced by the right-hand side of (2.1).

The LTS estimator can be defined as follows (Rousseeuw and van Driessen, 2000). Let ζ denote an h-subset of $(1,\dots,n)$ with associated least squares estimators

$$\hat\beta_\zeta = (\sum_{i\in\zeta} x_{in}x_{in}')^{-1}\sum_{i\in\zeta} x_{in}y_i \qquad \text{and} \qquad \hat\sigma_\zeta^2 = h^{-1}\sum_{i\in\zeta}(y_i - x_{in}'\hat\beta_\zeta)^2, \tag{2.2}$$

where $\sum_{i\in\zeta} x_{in}x_{in}'$ is assumed invertible for any ζ. Then, the LTS estimator and the associated scale estimator are given by

$$\hat\beta = \hat\beta_{\hat\zeta} \qquad \text{and} \qquad \hat\sigma^2 = \hat\sigma^2_{\hat\zeta} \qquad \text{where} \qquad \hat\zeta = \arg\min_\zeta \hat\sigma^2_\zeta. \tag{2.3}$$

That is, for a given number of "good" observations h, the LTS estimator finds the h-subsample with the smallest residual sum of squares.

2.2. The LTS Estimator in the ϵ-Contamination Model

The ϵ-contamination model of Huber (1964) has traditionally been the primary framework for analyzing the asymptotic properties of the LTS estimator. In this context, the errors, ε_i, are assumed to be i.i.d. with a common distribution $\mathsf{F} = (1-\epsilon)\Phi + \epsilon\mathsf{G}$ that mixes the standard normal distribution Φ with a contamination distribution G at a contamination level $0 \le \epsilon < 1$. Moreover, the errors are independent of the regressors.

In this stream of the literature, the asymptotic properties of the LTS estimator are derived by imposing conditions on the error distribution F, which in turn restricts the type of contamination distribution G. For instance, in an early paper, Butler (1982) studied the asymptotic properties of the LTS estimator for the location-scale case assuming a unimodal but not necessarily symmetric F. Rousseeuw (1985) and Croux and Rousseeuw (1992) described the asymptotic behavior of the LTS estimators for regression and scale, respectively. Víšek (2006) analyzed the linear regression case with symmetric F and compact parameter space. Using

Víšek (2006), Johansen and Nielsen (2016) analyzed the consistency of the scale estimator in the regression case.

More precisely, the above papers show that the asymptotic behavior of the regression and scale LTS estimators in this setting is as follows. Let $h = \lfloor \lambda n \rfloor$ so that λ denotes the proportion of "good" observations associated to h and $\lfloor . \rfloor$ the floor function. Suppose (ε_i, x_{in}) is i.i.d. with fourth moments. Let ε_i have a continuous, symmetric, unimodal distribution F with density f and be independent of x_{in} for which $\Sigma_x = \mathsf{E} x_{in} x_{in}'$. Then,

$$n^{1/2}(\hat{\beta} - \beta) \overset{\mathsf{D}}{\to} \mathsf{N}(0, \sigma^2 \Sigma_x^{-1} \eta_{\lambda,\mathsf{F}}), \qquad h^{-1} \sum_{i \in \hat{\zeta}} x_{in} x_{in}' \overset{\mathsf{P}}{\to} \Sigma_x, \qquad \hat{\sigma}^2 \overset{\mathsf{P}}{\to} \sigma^2 \varsigma^2_{\lambda,\mathsf{F}},$$

with efficiency factor $\eta_{\lambda,\mathsf{F}}$ and consistency factor $\varsigma^2_{\lambda,\mathsf{F}}$ given by

$$\eta_{\lambda,\mathsf{F}} = \frac{\int_{-c}^{c} x^2 d\mathsf{F}(x)}{\{\lambda - 2c\mathsf{f}(c)\}^2}, \qquad \varsigma^2_{\lambda,\mathsf{F}} = \frac{\int_{-c}^{c} x^2 d\mathsf{F}(x)}{\lambda}, \qquad c = \mathsf{F}^{-1}\{(1+\lambda)/2\}.$$

Both, $\eta_{\lambda,\mathsf{F}}$ and $\varsigma^2_{\lambda,\mathsf{F}}$, depend on the unknown distribution F. Inference, then, depends on these nuisance parameters. The state-of-the-art in this context is to proceed assuming that the errors, ε_i, are normal so that $\epsilon = 0$. This prevalent approach seems contrary to the original intention of using robust estimation to deal with outliers. Nonetheless, it is customary in the literature to follow this method and, therefore, we will take it as a framework for comparison with our approach below. We refer to this approach as standard LTS or in short SLTS inference. Under normal errors, $\int_{-c}^{c} x^2 d\Phi(x) = \lambda - 2c\phi(c)$, so that $\eta_{\lambda,\Phi} = 1/\{\lambda - 2c\phi(c)\}$. Therefore, under normality, $\lambda \eta_{\lambda,\Phi} = 1/\varsigma^2_{\lambda,\Phi}$. The SLTS inferential approach uses, for $h = \lambda n$, the estimated asymptotic variance

$$\widehat{\mathsf{Var}}_{SLTS}(\hat{\beta}) = \frac{\hat{\sigma}^2}{n \varsigma^2_{\lambda,\Phi}} \Big(h^{-1} \sum_{i \in \hat{\zeta}} x_{in} x_{in}' \Big)^{-1} \eta_{\lambda,\Phi} = \hat{\sigma}^2 \Big(\sum_{i \in \hat{\zeta}} x_{in} x_{in}' \Big)^{-1} \Big(\frac{1}{\varsigma^2_{\lambda,\Phi}} \Big)^2. \qquad \textbf{(2.4)}$$

2.3. The LTS Model

Departing from the traditional ϵ-contamination approach, Berenguer-Rico et al. (2023) proposed the following model in which the LTS estimator is maximum likelihood.

Model 1. (The LTS model). Let $y_i = x_{in}'\beta + \sigma\varepsilon_i$ for $i = 1, ..., n$. We condition on the random regressors $x_{1n}, \ldots, x_{nn}$. Let $h \leq n$ be given and ζ be a non-random, given set with h elements from $1, \ldots, n$.

For $i \in \zeta$, let ε_i be i.i.d. $\mathsf{N}(0,1)$ distributed.

For $j \notin \zeta$, let ξ_j be independent with distribution functions $\mathsf{G}_j(z)$ for $z \in \mathbb{R}$, where G_j is continuous at 0, but may depend on x_{jn}. The "outlier" errors are defined, for $j \notin \zeta$, by

$$\varepsilon_j = (\max_{i \in \zeta} \varepsilon_i + \xi_j) 1_{(\xi_j > 0)} + (\min_{i \in \zeta} \varepsilon_i + \xi_j) 1_{(\xi_j < 0)}. \qquad \textbf{(2.5)}$$

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

The parameters are $\beta \in \mathbb{R}^{\dim x}$, $\sigma > 0$, ζ which is any h-subset of $i = 1, \ldots, n$ and G_j which are any $n-h$ arbitrary conditional distributions on $\mathbb{R}$, that are continuous at 0.

The LTS model allows for many types of "outliers". Its defining feature is that the "outlier" errors are outside the realized range of the "good" errors and are characterized by an un-specified distribution $\mathsf{G}_j(z)$. Given the semi-parametric nature of the model, Berenguer-Rico et al. (2023) used the ϵ-likelihood concept of Scholz (1980) to show that the LTS estimator is maximum likelihood in the LTS model.

Unlike the ϵ-contamination model described above, the LTS model does not have an i.i.d. structure. The "outlier" errors are beyond the realized range of the "good" errors. Hence, there is dependence in the LTS model. Moreover, the "outlier" errors, ε_j, have distribution function G_j, which can vary with j and depend on x_{jn}. Hence, there is heterogeneity in the LTS model. Thus, the probabilistic structure of the contaminated observations in the ϵ-contamination model and the LTS model are different. And, as we will see below, they will bring about very different properties of the LTS estimator.

The LTS model can generate a great variety of contamination schemes. Apart from taking values outside the realized range of the "good" errors, the "outlier" errors are very much unrestricted and could even depend on the regressors. When showing that the LTS estimator is maximum likelihood in the LTS model, the regressors were conditioned upon. Hence, the regressors in the LTS model can be generated in multiple ways: fixed or random, i.i.d. or time series, with or without "outliers", etc. We illustrate some of the different types of "outliers" that the LTS model can generate in a series of simulated data. Specifically, we consider the linear model $y_i = \beta_0 + \beta_1 z_i + \sigma\varepsilon_i$, for $i = 1, ..., 100$, and generate six different data generating processes (DGPs). All DGPs have $\beta_0 = \beta_1 = 0$ and $\sigma = 1$. In all cases, the "good" errors are i.i.d. $\mathsf{N}(0,1)$ and the "good" regressors are i.i.d. uniform on $[-10, 10]$, denoted $\mathsf{U}[-10, 10]$.

Figure 1 gives scatter plots of y_i against z_i for the six different DGPs with "good" and "outlying" observations marked with crosses and bullets, respectively. The dotted lines indicate the maximal absolute "good" errors. Also shown are the full sample OLS (solid line, grey) and LTS with $h = 80$ (dashed line, grey) fits.

DGP 1 has no contamination and is illustrated in panel (a). All errors are i.i.d. $\mathsf{N}(0,1)$ and all regressors are i.i.d. $\mathsf{U}[-10, 10]$. The OLS and LTS (with wrong $h = 80$) lines are both very close to the zero line.

DGPs 2–6, illustrated in panels (b)–(f), have contamination of LTS type. In all cases, there are $h = 80$ "good" observations (crosses) and $n - h = 20$ "outliers" (bullets).

DGPs 2–3 are illustrated in panels (b) and (c), respectively. In both cases, there is only contamination in the error term and the regressors are all i.i.d. $\mathsf{U}[-10, 10]$. The error term has LTS-type contamination of the form (2.5) where $\xi_j - \nu^+ 1_{(\xi_j > 0)} + \nu^- 1_{(\xi_j < 0)}$ is i.i.d. normal $\mathsf{N}(0,1)$. The constants ν^+ and ν^- separate "good" and

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

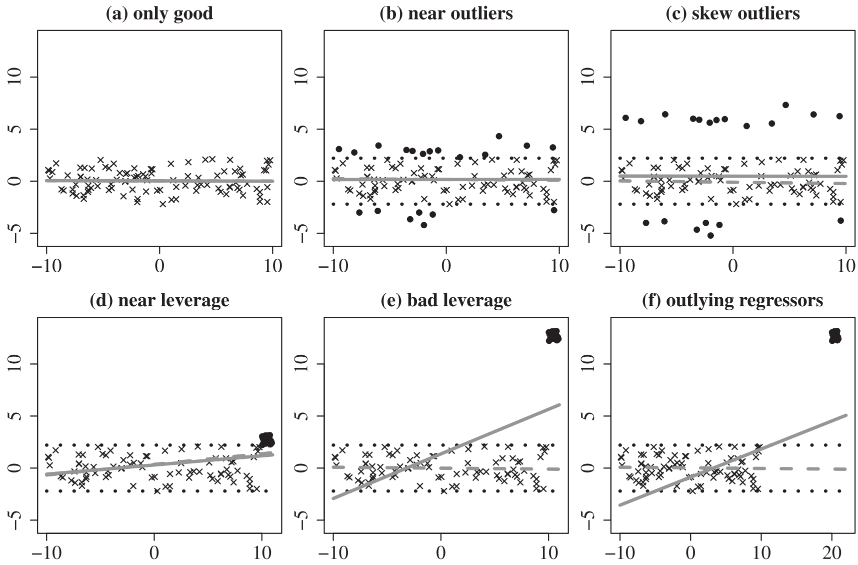

FIGURE 1. Scatter plots for six different DGPs from the LTS model, $n = 100$. Points are "good" (cross) and "outliers" (bullet). Lines are max absolute "good" error (dots), LTS fit (dashed, grey), OLS fit (solid, grey).

"outlying" errors. DGP 2 has $\nu^+ = \nu^- = 0$ (near "outliers") and DGP 3 has $\nu^+ = 3$, $\nu^- = 1$ (skew "outliers"). In DGP 2, with near "outliers", the OLS and LTS lines are both very close to the zero line. In DGP 3, with separation and skew "outliers", the OLS estimator has slope close to zero but the intercept is off. In contrast, the LTS line is close to the zero line.

DGPs 4–6, depicted in panels (d)–(f), have contamination in both errors and regressors. Both "outlier" errors and regressors are of LTS type. Specifically, the "outlier" errors are of the form (2.5) with $\xi_j = u_j + c$ and u_j i.i.d.U$[0,1]$. The "outlier" regressors are $x_j = 10 + e_j + d$, where e_j are i.i.d.U$[0,1]$ and independent of u_j, although in the theory below they do not need to be independent. Note that the "good" regressors are U$[-10,10]$, hence, the "outlier" regressors are also outside of the range of the "good" regressors. DGP 4 has $c = d = 0$, generating a near leverage effect. Both OLS and LTS lines seem to be attracted to the near leverage "outlier" points in this small sample of size $n = 100$. DGP 5 has $c = 10$ and $d = 0$, which generates bad leverage points. The OLS line is clearly attracted to the bad leverage "outliers" while the LTS line is close to the zero line. DGP 6 has $c = 10$, $d = n^{1/2}$, so that the "outlier" regressors are $x_j = 20 + e_j$ and therefore outlying themselves. As in DGP 5, the OLS line is attracted by the "outliers", while the LTS line remains close to the zero line. The theory that follows will cover DGPs 1-5. DGP 6, which has $d = n^{1/2}$, and therefore regressor "outliers" that diverge with the sample size, has been chosen as an extreme example in which the LTS estimator

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

fails asymptotically. This is in line with the recommendation (Rousseeuw, 1994) of looking for “outliers” in the regressors before using the LTS estimator. See the running example in Section 3 and simulations below for more details.

A feature of the LTS model is that “outlier” values are relative to a given sample. The “outlier” errors are always beyond the realized range of the “good” normal errors. The “good” normal errors have unbounded support, therefore, as the sample size increases, the “outlier” errors are more and more extreme. This implies that values that are “outliers” in one sample will not be so in larger samples.

One will notice that the LTS model is not a mixture model with an i.i.d. structure, for which the index set for the “outliers” would have been random. Rather, in the LTS model “outlier” errors are conditional on the “good” errors and, therefore, are not i.i.d. Moreover, the index set ζ is fixed and a parameter of the model. The LTS model is therefore different from the Huber (1964) ϵ-contamination model. The LTS model, with its likelihood underpinning, brings a new alternative way to probabilistically study and incorporate “outliers” in statistical analysis that can be fruitful. For instance, as we show below, the LTS model approach delivers nuisance parameter free asymptotic inference in the LTS regression context, which can be useful in practice when conducting inference using the LTS estimator.

As illustrated by Figure 1, the flexible and semi-parameteric nature of the LTS model allows for a great variety of contamination schemes. Many other configurations can be generated as well. Figure 1 has contamination types that are graphically similar to “outliers” considered previously in the literature. Here, the “outliers” are generated systematically through the LTS model with its probabilistic structure. This provides an analytic framework for the theoretical study of the LTS estimator (and potentially other estimators) under general contamination.

3. LTS ASYMPTOTICS

Berenguer-Rico et al. (2023) showed that the LTS estimator is maximum likelihood in the LTS regression model with $y_i = x_{in}'\beta + \sigma\varepsilon_i$. Asymptotics for the LTS estimator in the LTS model were analyzed by Berenguer-Rico et al. (2023) but only in the location-scale case, $y_i = \mu + \sigma\varepsilon_i$, where a linear search suffices. Analyzing the asymptotic properties of the LTS estimator in the regression case, $y_i = x_{in}'\beta + \sigma\varepsilon_i$, which uses a binomial search, is a non-trivial extension, requiring new asymptotic tools as provided here.

Hence, the following theory uses the linear regression equation $y_i = x_{in}'\beta + \sigma\varepsilon_i$ in (2.1). The LTS model with the maximum likelihood property as described in Section 2.3 is taken as a benchmark and it will be relaxed in a series of assumptions. For instance, separation between “good” and “outlier” errors will be kept but in a less stringent way so that some overlap is permitted. Moreover, the “good” errors do not need to be normal. Instead, conditions on their extreme and intermediate values will be assumed. In this way, other commonly used distributions will be allowed. Also, instead of conditioning on the regressors, we will introduce regularity conditions for the regressors allowing them to be random. All in all, the asymptotic theory derived below does not assume the LTS model as such. While the stated assumptions capture its essence, they are more general and allow for a much broader class of models. This is in line with OLS theory. OLS is maximum likelihood in the normal model but usually its asymptotic properties are derived under a relaxed set of assumptions. We proceed similarly here.

For the asymptotic results in this section, the linear regression equation (2.1) is used, that is $y_i = x_{in}'\beta + \sigma\varepsilon_i$ for $i = 1,\dots,n$. Let $h \le n$ be the number of "good" observations, which we assume to be known throughout. In the asymptotics, we consider increasing values of h,n. Assumptions for the errors ε_i and the regressors x_{in} are given as we progress. Examples of permitted regressors are discussed in Section 4. We let ζ_n be a deterministic sequence of sets of true indices of "good" observations and ζ be a generic h-set of indices from $i = 1,...,n$. Let $\#\zeta$ denote the count of elements in the set ζ.

To illustrate the assumptions below, we will use the LTS model with leverage (DGPs 4-6 above) as a running example.

Running Example. Suppose $y_i = x_{in}'\beta + \sigma\varepsilon_i$ where $x_{in}' = (1, z_{in})$ is bivariate. Let $1/2 < \lambda \le 1$. Further, let $\lfloor . \rfloor$ denote the floor function and define $h = \lfloor n\lambda \rfloor$ to be the number of "good" observations. Let ζ_n be the set of indices of "good" observations.

For $i \in \zeta_n$, ε_i are i.i.d.$\mathsf{N}(0,1)$ and z_{in} are i.i.d.$\mathsf{U}[-10,10]$.

For $j \notin \zeta_n$, the "outlier" errors and regressors are defined by $\varepsilon_j = \max_{i\in\zeta_n}\varepsilon_i + \xi_j$ for $\xi_j = u_j + c$ and $z_{jn} = 10 + e_j + d$, respectively, where u_j and e_j are i.i.d.$\mathsf{U}[0,1]$. We assume here that u_j and e_j are independent but the theory allows dependence. In DGP 4, $c = d = 0$. In DGP 5, $c = 10, d = 0$. In DGP 6, $c = 10, d = n^{1/2}$. As illustrated in Figure 1, all these data generating processes produce data with leverage points.

3.1. Boundedness

A boundedness result is presented for the LTS estimator for the linear regression equation (2.1), that is $y_i = x_{in}'\beta + \sigma\varepsilon_i$, under assumptions to the second sample moment for the "good" errors and to the frequency of small regressors.

Assumption 3.1. Suppose

(i) **Frequency of "good" observations**: $h/n \to \lambda$ where $\lambda > 1/2$.
(ii) **"Good" errors**: $h^{-1}\sum_{i\in\zeta_n}\varepsilon_i^2 = \mathrm{O}_{\mathsf{P}}(1)$.
(iii) **Frequency of small regressors**: Define

$$F_{nh}(a) = \max_{\zeta:\#\zeta=h}\sup_{\delta:|\delta|=1} h^{-1}\sum_{i\in\zeta} 1_{(|x_{in}'\delta|\le a)}. \tag{3.1}$$

Let ξ satisfy $0 < \xi < 2 - \lambda^{-1}$ and suppose

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

$$\lim_{(a,n)\to(0,\infty)} \mathsf{P}\{F_{nh}(a) > \xi\} = 0, \tag{3.2}$$

that is $\forall\epsilon > 0$, $\exists(a_0,n_0) > 0$: $\forall a \le a_0, n \ge n_0$ then $\mathsf{P}\{F_{nh}(a) > \xi\} < \epsilon$.

Remark 3.1. Assumption 3.1(*ii*) implies that $\hat\sigma^2$ is bounded in probability. Indeed, since $\hat\sigma^2$ is a minimizer then $\hat\sigma^2 \le \hat\sigma^2_{\zeta_n}$ where $\hat\sigma^2_{\zeta_n}/\sigma^2 \le h^{-1}\sum_{i\in\zeta_n}\varepsilon_i^2$ by the model equation (2.1).

Remark 3.2. Assumption 3.1(*iii*) implies that $\hat\Sigma_\zeta = h^{-1}\sum_{i\in\zeta} x_{in}x'_{in}$ is positive definite in large samples for all ζ as required in (2.2), see Section 4.1 below. It covers a wide range of regressors – examples are given in Section 4.

Remark 3.3. The boundedness result in this section, its condition for the frequency of small regressors and its proof are inspired by the analysis of M-estimators in Johansen and Nielsen (2019). A major difference is that the objective functions for M-estimators and the LTS estimator have a different structure. M-estimators minimize $\sum_{i=1}^n \rho(y_i - x'_{in}\beta)$ for a criterion function ρ, that may be bounded and the theory is formulated for a general ρ. The LTS objective function is built around a quadratic criterion function that sums over a set of "good" observations that is to be estimated. When $h = n$, the LTS estimator is therefore an OLS estimator as well as an M-estimator. Moreover, when $h = n$, the $F_{nn}(a)$ function in (3.1) is equal to the quantity $F_n(a)$ in Johansen and Nielsen (2019) which was used to prove boundedness of M-estimators. It is interesting to note that the $F_n(a)$ function is related to quantities found in previous papers on M-estimators (Chen and Wu, 1988) and on S-estimators (Davies, 1990).

Running Example. Assumption 3.1(*i*) holds in this example since $1/2 < \lambda \le 1$.

Assumption 3.1(*ii*) holds since the "good" errors ε_i for $i \in \zeta_n$ are i.i.d.N$(0,1)$, so that $h^{-1}\sum_{i\in\zeta_n}\varepsilon_i^2$ is χ^2_h/h distributed and bounded in probability.

Assumption 3.1(*iii*) holds as follows. We can bound $F_{nh}(a) \le F_{hh}(a) + (n-h)/h$ with the convention $F_{hh}(a) = \sup_{\delta:|\delta|=1} h^{-1}\sum_{i\in\zeta_n} 1_{(|x'_{in}\delta|\le a)}$. This is proved in (4.2) below. Notice that the bound only involves the "good" regressors. For $i \in \zeta_n$, we have $x'_{in} = (1, z_{in})$ where z_{in} is i.i.d. with bounded, continuous density in this example. We get $F_{hh}(a) = \mathrm{o}_\mathsf{P}(1)$ (Johansen and Nielsen, 2019, Thm. 3.3). Moreover, $(n-h)/h \to \lambda^{-1} - 1$ as $h/n \to \lambda$. Thus, we can bound $F_{nh}(a) \le \lambda^{-1} - 1 + \mathrm{o}_\mathsf{P}(1)$. We have that $\lambda^{-1} - 1 < 2 - \lambda^{-1}$, whenever $\lambda > 2/3$. This leaves space for choosing a ξ so that Assumption 3.1(*iii*) is satisfied.

Under Assumption 3.1, the first result bounds the difference between the LTS estimator and the infeasible OLS estimator, $\hat\beta_{\zeta_n}$, on the unknown set of "good" observations, whose indices are given in the deterministic set ζ_n. The asymptotic theory of the OLS estimator $\hat\beta_{\zeta_n}$ is of course widely studied. We note that the LTS estimator may not be unique, so we establish a uniform bound over the sets $\mathcal{M}_n$ of minimizers ζ of $\hat\sigma^2_\zeta$.

THEOREM 3.1. *Suppose Assumption 3.1. Let $\mathcal{M}_n$ denote the set of minimizers ζ of $\hat\sigma^2_\zeta$. Then,* $\max_{\zeta\in\mathcal{M}_n}|\hat\beta_\zeta - \hat\beta_{\zeta_n}| = \mathrm{O}_\mathsf{P}(1)$.

The boundedness of the LTS estimator derived in Theorem 3.1, holds under very mild assumptions on the "good" errors and the frequency of small regressors. No other structure than Assumption 3.1 is imposed. In particular, the defining feature of the LTS Model of placing "outlier" errors outside the realized range of "good" errors is not used. This means that the result can be used in a range of situations, like the LTS or ϵ-contamination models described above.

Remark 3.4. Theorem 3.1 also provides some insights on the behavior of the LTS estimator when the wrong h is chosen. Suppose that Assumption 3.1 holds for $h_\circ$ with $h_\circ/n \to \lambda_\circ$ and the LTS procedure is used with $h < h_\circ$ observations so that $h/n \to \lambda$ and $1/2 < \lambda < \lambda_\circ$. We study whether Assumption 3.1 will be satisfied for the sequence $h < h^\circ$. First, Assumption 3.1(i) holds by construction since we choose h so that $h/n \to \lambda$ and $1/2 < \lambda < \lambda_\circ$.

Second, if Assumption 3.1(ii) holds for $h_\circ$, then there exists a set $\zeta_\circ$ in which $h_\circ^{-1}\sum_{i\in\zeta_\circ}\varepsilon_i^2 = \mathrm{O}_\mathsf{P}(1)$. Let $\zeta_n \subset \zeta_\circ$ be a subset of size $h < h_\circ$ of the indices included in $\zeta_\circ$. Then, $\sum_{i\in\zeta_n}\varepsilon_i^2 \le \sum_{i\in\zeta_\circ}\varepsilon_i^2 = \mathrm{O}_\mathsf{P}(h_\circ)$. Since $h/h_\circ \to \lambda/\lambda_\circ$, then $h^{-1}\sum_{i\in\zeta_n}\varepsilon_i^2 = \mathrm{O}_\mathsf{P}(1)$. Third, suppose $F_{nn}(a) = \mathrm{o}_\mathsf{P}(1)$, which is valid in the Examples 4.1–4.4 below. Then for any index set ζ we can bound $\sum_{i\in\zeta}1_{(\cdot)} \le \sum_{i=1}^{n}1_{(\cdot)}$. In particular, we can bound $F_{nh_\circ}(a) \le (n/h_\circ)F_{nn}(a) = \mathrm{o}_\mathsf{P}(1)$ and $F_{nh}(a) \le (n/h)F_{nn}(a) = \mathrm{o}_\mathsf{P}(1)$ so that Assumption 3.1(*iii*) holds for $h_\circ$ as well as for the h chosen for LTS estimation.

If instead, the LTS procedure is used with $h > h_\circ$ observations, then Assumption 3.1 may or may not apply for the h sequence. Specifically, when $h > h_\circ$, it would be possible to construct cases in which the LTS estimator is bounded and cases in which the LTS estimator is unbounded. Notice that when $h > h_\circ$, any choice of the h-set ζ_n in Assumption 3.1 must include both "good" observations and "outliers" due to the constraint that $h > n/2$. If the "outlier" errors are diverging, then Assumption 3.1(*ii*) for the h sequence will fail and this could translate into unboundedness of LTS.

3.2. Consistent Selection of "Good" Observations

Next, we show that the proportion of wrongly classified observations vanishes. The convergence rate is improved subsequently. We note that $\#(\zeta \cap \zeta_n)$ is the number of "good" observations that are correctly selected in ζ. The numbers of wrongly classified "good" observations and wrongly classified "outliers" satisfy $\#(\zeta^c \cap \zeta_n) = \#(\zeta \cap \zeta_n^c)$, since $h = \#(\zeta^c \cap \zeta_n) + \#(\zeta \cap \zeta_n)$ and $h = \#(\zeta \cap \zeta_n^c) + \#(\zeta \cap \zeta_n)$. The proportion of wrong classifications is then $\#(\zeta \cap \zeta_n^c)/h$. Let $\|m\|$ denote the spectral norm of a matrix m.

Assumption 3.2. Suppose

(i) **Regressors**: $\|\sum_{i=1}^{n}x_{in}x_{in}'\| = \mathrm{O}_\mathsf{P}(n)$.

(ii) **Infeasible OLS estimator**: $\hat{\beta}_{\zeta_n} = \mathrm{O}_\mathsf{P}(1)$.

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

Remark 3.5. Assumption 3.2(i) is a mild assumption to the regressors. It allows for "outlier" regressors but these should not be divergent. This is consistent with the recommendation of Rousseeuw (1994) to start an LTS analysis by detecting "outliers" among the regressors. See the running example below and Section 4 for other specific examples.

Running Example. Assumption 3.2(i) holds for DGPs 4-5 as follows. The "good" regressors are i.i.d.$\mathsf{U}[-10,10]$, so that $z_{in}^2 \leq 100$ *a.s.* whence $h^{-1}\sum_{i\in\zeta_n} z_{in}^2 \leq 100$ *a.s.* Similarly, the "outlier" regressors are i.i.d.$\mathsf{U}[10+d, 11+d]$, so that $(n-h)^{-1}\sum_{i\notin\zeta_n} z_{in}^2 \leq (11+d)^2$ *a.s.* These findings can be combined to show that $\|\sum_{i=1}^n x_{in}x_{in}'\| = \mathsf{O}_\mathsf{P}(n)$.

Assumption 3.2(i) fails for DGP 6. The "outlier" regressors are i.i.d.$\mathsf{U}[10+d, 11+d]$ with $d = n^{1/2}$, so that $(n-h)^{-1}\sum_{i\notin\zeta_n} z_{in}^2 \geq (10+d)^2 \to \infty$ *a.s.* Assumption 3.2(ii) holds as follows. Since the "good" errors are normal, it suffices that $\sum_{i\in\zeta_n}^n x_{in}x_{in}'$ is bounded from below. This holds when $x_{in} = (1, z_{in})'$ and z_{in} are i.i.d.$\mathsf{U}[-10,10]$.

Theorem 3.2. *Suppose Assumptions 3.1 and 3.2. Let $\mathcal{M}_n$ denote the set of minimizers ζ of $\hat{\sigma}_\zeta^2$. Then,* $\max_{\zeta\in\mathcal{M}_n}\#(\zeta\cap\zeta_n^c)/h = \mathsf{O}_\mathsf{P}(1/\min_{j\notin\zeta_n}\varepsilon_j^2)$.

It is worth noting that the LTS model as defined in Section 2.3 is not used in proving Theorem 3.2. The proof of Theorem 3.2 is derived under Assumptions 3.1 and 3.2 only.

Theorem 3.2 provides a consistency result whenever the smallest "outlier" error squared, $\min_{j\notin\zeta_n}\varepsilon_j^2$, diverges. We then have that the proportion of wrong classifications vanishes in that $\#(\zeta\cap\zeta_n^c)/h = \mathsf{o}_\mathsf{P}(1)$. Since $\#(\zeta\cap\zeta_n)+\#(\zeta\cap\zeta_n^c) = h$, we also get that the proportion of correctly classified "good" observations goes to unity, that is $\#(\zeta\cap\zeta_n)/h = 1+\mathsf{o}_\mathsf{P}(1)$.

Running Example. Theorem 3.2 provides a consistency result if $\min_{j\notin\zeta_n}\varepsilon_j^2$ diverges. We show this is the case. Recall that the "outliers" are $\varepsilon_j = \max_{i\in\zeta_n}\varepsilon_i + \xi_j$ where $\xi_j = u_j + c > 0$ *a.s.* as $c = 0$ or $c = 10$ and u_j are i.i.d.$\mathsf{U}[0,1]$. The "good" errors are i.i.d. standard normal, so that $\max_{i\in\zeta_n}\varepsilon_i/\sqrt{2\log h} \to 1$ *a.s.*, see Example B.1 below. We also have that $\min_{j\notin\zeta_n} u_j \to 0$ *a.s.* whence $\min_{j\notin\zeta_n}\xi_j \to c$. In combination, we get, for finite c, that $\min_{j\notin\zeta_n}\varepsilon_j/\sqrt{2\log h} \to 1$ *a.s.* Hence, $\min_{j\notin\zeta_n}\varepsilon_j^2$ diverges, so that Theorem 3.2 then shows that $\max_{\zeta\in\mathcal{M}_n}\#(\zeta\cap\zeta_n^c)/h = \mathsf{O}_\mathsf{P}(1/\log h)$.

In the running example, we have that the smallest "outlier" error squared, $\min_{j\notin\zeta_n}\varepsilon_j^2$, is no less than the largest "good" error, $\max_{i\in\zeta_n}\varepsilon_i^2$. The latter diverges at a logarithmic rate because the "good" errors are normal. In general, $\max_{i\in\zeta_n}\varepsilon_i^2$ diverges when the "good" errors are i.i.d. with unbounded support. However, if $\min_{j\notin\zeta_n}\varepsilon_j^2$ is bounded in probability, then Theorem 3.2 says that the wrong classification rate is bounded in probability. This is rather uninformative as the wrong classification rate is at most unity by construction. In contrast, when

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

$\min_{j \notin \zeta_n} \varepsilon_j^2$ diverges faster than h, then Theorem 3.2 says that we will get perfect classification $\#(\zeta \cap \zeta_n^c) = \mathrm{o_P}(1)$. A variation of the latter case is discussed in Appendix D.

3.3. Improving the Rate of Consistency

Theorem 3.2 gave conditions under which $\#(\zeta \cap \zeta_n^c)/h = \mathrm{O_P}(1/\min_{j \notin \zeta_n} \varepsilon_j^2)$, which will typically be a slow rate. Here, we improve the consistency rate by making assumptions to the intermediate extreme values of the "good" errors and regressors.

Assumption 3.3. Let $m_n^2 = \min\{(\min_{i \in \zeta_n} \varepsilon_i)^2, (\max_{i \in \zeta_n} \varepsilon_i)^2\}$. Suppose

(i) **"Good" errors**: ε_i for $i \in \zeta_n$ satisfy
  (a) $1/m_n^2 = \mathrm{o_P}(1)$;
  (b) $\max_{i \in \zeta_n} \varepsilon_i^2 / m_n^2 = \mathrm{O_P}(1)$;
  (c) Let ε_i^2 for $i \in \zeta_n$ have order statistics $\psi_1 \leq \cdots \leq \psi_h$. Then, the intermediate extreme values satisfy $\forall 0 < \rho < 1$, $\exists C_\rho < 1$: $\psi_{h - \lfloor h^\rho \rfloor}/m_n^2 \leq C_\rho + \mathrm{o_P}(1)$;
  (d) Extremes are of polynomial order: $m_n^2 = \mathrm{O_P}(n^\eta) L(n)$ for some $0 \leq \eta < 1/2$ and where $L(n)$ is a slowly varying function, that is $L(an)/L(n) \to 1$ as $n \to \infty$ for any $a > 0$.

(ii) **"Outlier" errors**: $\min_{j \notin \zeta_n} \varepsilon_j^2 \geq m_n^2 \{1 + \mathrm{o_P}(1)\}$.

(iii) **Regressors**: Let $|x_{in}|$ have order statistics $x_{(1)} \leq \cdots \leq x_{(n)}$ satisfying either
  (a) $x_{(n)} = \mathrm{O_P}(1)$; or
  (b) $x_{(n)}^2 = \mathrm{O_P}(m_n^2)$ and $\forall 0 < \delta < 1$, $\exists 0 < r < 1 - \eta$: $x_{(n - \lfloor n^r \rfloor)}^2 / x_{(n)}^2 \leq \delta\{1 + \mathrm{o_P}(1)\}$.

(iv) **Infeasible OLS estimator**: $(\hat{\beta}_{\zeta_n} - \beta)' (\sum_{i \in \zeta_n} x_{in} x_{in}') (\hat{\beta}_{\zeta_n} - \beta) = \mathrm{O_P}(1)$.

Remark 3.6. Assumption 3.3(i) concerns the tail behavior of the "good" errors. The extreme tail conditions in (a, b, d) can be assessed using the multiplicative strong law of large numbers (Galambos, 1978, Thm. 4.4.4). The intermediate tail condition (c) can be assessed by modifying Chibisov (1964, Lem. 1). See Appendix B for details.

Example 3.1. Assumption 3.3(i) holds in the following cases. Appendix B gives details.

(i) Normal distribution with $m_n^2 / 2 \log h \to 1$ a.s. and $C_\rho = 1 - \rho$;
(ii) Laplace distribution with $m_n / \log h \to 1$ a.s. and $C_\rho = (1 - \rho)^2$;
(iii) Double geometric distribution with $m_n / \log h \to 1$ a.s., $C_\rho = (1 - \rho)^2$;
(iv) t_d distribution with $d > 4$ degrees of freedom. In this case, $m_n / h^{1/d}$ converges in distribution and any choice of C_ρ function suffices.

Remark 3.7. Assumption 3.3(ii) relaxes the defining feature of the LTS model that "outlier" errors are more extreme than "good" errors by allowing some overlap between "outlier" errors and "good" errors. As an example of such overlap, suppose the "good" errors are standard normal. In that case, $\max_{i \in \zeta_n} \varepsilon_i^2 / 2 \log h \to 1$ *a.s.*

If the smallest "outlier" error is $\min_{j\notin\zeta_n}\varepsilon_j = \sqrt{2\log h} - 1$, then (a) $\mathsf{P}(\min_{j\notin\zeta_n}\varepsilon_j < \max_{i\in\zeta_n}\varepsilon_i) \to 1$, so there is overlap with high probability, and (b) $\min_{j\notin\zeta_n}\varepsilon_j^2 = m_n^2\{1 + \mathrm{o}_{\mathsf{P}}(1)\}$, so Assumption 3.3(*ii*) is satisfied. See Example B.5 in Appendix B for details.

Remark 3.8. Assumption 3.3(*iii*) restricts the regressors' tails. Essentially, the regressors cannot have thicker tails than the "good" errors. Even so, the assumption allows a large variety of "good" and "outlier" regressors. Examples follow in Section 4.

Remark 3.9. Assumption 3.3(*iv*) restricts the joint distribution of "good" errors and regressors. Suppose (x_i, ε_i) are i.i.d. for $i \in \zeta_n$. Then, we will need that $\mathsf{E}x_i\varepsilon_i = 0$ and a Central Limit Theorem to achieve boundedness of the normalized infeasible OLS estimator. In particular, Assumption 3.3(*iv*) fails if $\mathsf{E}\varepsilon_i = 0$, but x_i and ε_i are correlated.

Remark 3.10. Assumption 3.3 is very flexible in regards to dependence between "outlier" errors and regressors. Neither $\mathsf{E}\varepsilon_j = 0$ nor $\mathsf{E}(\varepsilon_j \mid x_j) = 0$ is imposed. This is key in allowing for leverage effects.

Running Example. Assumption 3.3(*i*) holds as follows. The "good" errors are normal and as noted in Example 3.1 all conditions (*a*)–(*d*) are satisfied. In particular, $m_n^2/(2\log h) \to 1$ *a.s.* As the logarithm is slowly varying we find that $m_n^2 = \mathrm{O}_{\mathsf{P}}(n^{\eta})L(n)$ with $\eta = 0$.

Assumption 3.3(*ii*) holds as follows. Since $\varepsilon_j = \max_{i\in\zeta_n}\varepsilon_i + \xi_j$ where $\xi_j \geq 0$ *a.s.*, we get that $\min_{j\notin\zeta_n}\varepsilon_j^2 \geq \max_{i\in\zeta_n}\varepsilon_i^2 \geq m_n^2$.

Assumption 3.3(*iii, a*) holds for DGPs 4-5 as follows. The "good" regressors z_{in} are i.i.d.$\mathsf{U}[-10, 10]$. The "outlier" regressors are $z_{jn} = 10 + e_j + d$ where e_j are i.i.d.$\mathsf{U}[0, 1]$ with $d = 0$ in DGPs 4 and 5. Thus, the regressors are bounded as required.

Assumption 3.3(*iii*) fails for DGP 6. The "good" regressors z_{in} are i.i.d.$\mathsf{U}[-10, 10]$. The "outlier" regressors are $z_{jn} = 10 + e_j + d$ where e_j are i.i.d.$\mathsf{U}[0, 1]$ and $d = \sqrt{n}$. Since $\max_{j\notin\zeta_n} e_j \to 1$ *a.s.*, then $x_{(n)}/\sqrt{n} \to 1$ *a.s.* As $x_{(n)}$ diverges at a $\sqrt{n}$ rate it is neither bounded nor bounded by the logarithmically growing m_n^2.

Assumption 3.3(*iv*) holds as the standardized infeasible least squares estimator based on the "good" normal observations is standard normal as remarked above.

THEOREM 3.3. *Suppose Assumptions 3.1–3.3. Let $\mathcal{M}_n$ denote the set of minimizers ζ of $\hat{\sigma}_\zeta^2$. Then, for all* $0 < \theta < 1$, *it holds* $\max_{\zeta\in\mathcal{M}_n}\#(\zeta \cap \zeta_n^c)/h = \mathrm{O}_{\mathsf{P}}(h^{\theta-1})$.

It is worth mentioning again that the LTS model as defined in Section 2.3 is not used in proving Theorem 3.3. The proof of Theorem 3.3 is derived under Assumptions 3.1, 3.2, and 3.3 only. Under these assumptions the consistency rate is improved to a polynomial rate, $h^{\theta-1}$. This rate is used in the next section to derive the main result.

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

Remark 3.11. Theorem 3.3 shows that the proportion of misclassified observations vanishes under Assumption 3.3 which allows near "outliers" and even a slight overlap of "outlier" and "good" errors. We suspect this is a common situation in practice. The stronger result of perfect classification seems to require that "outliers" are far away from the regression line. This situation is analyzed in Appendix D.

More specifically, in Appendix D, we relax Assumption 3.3(i,a) which requires "good" errors with unbounded support. Allowing for "good" errors that can be either bounded or unbounded while imposing conditions on the growth of the "outlier" errors gives sharper results on the asymptotic properties of the LTS estimator. Perfect classification can be obtained. However, near "outliers", which can be common in practice, are not allowed in that case, see Appendix D for details. We also note that for bounded "good" errors, other estimators than the LTS estimator might be more desirable. For instance, when the "good" errors are uniform, then the Least Median Squares (LMS) estimator of Rousseeuw (1984) is maximum likelihood and LMS is n-consistent in the location scale model, see (Berenguer-Rico et al., 2023, supplement).

3.4. Main Result

Next, we show that the asymptotic distribution of the normalized LTS estimator coincides with that of the normalized infeasible OLS estimator on the "good" observations.

The result below involves a joint diagonalization. As $M = \sum_{i\in\zeta} x_{in}x'_{in}$ and $N = \sum_{i\in\zeta_n} x_{in}x'_{in}$ are symmetric and positive definite, there exists an invertible matrix S and a diagonal matrix Λ so that $N = SS'$ and $M = S(I_{\dim x} + \Lambda)S'$ (Johansen, 1995, Lem. A.5). We define the right square roots $N^{1/2} = S'$ and $M^{1/2} = (I_{\dim x} + \Lambda)^{1/2}S'$. The elements λ of Λ solve the equation $\det\{(1+\lambda)N - M\} = 0$ so that $1+\lambda > 0$ with corresponding eigenvectors v, such that $(1+\lambda)Nv = Mv$. In matrix notation, we have $V'NV = I_{\dim x}$ and $V'MV = I_{\dim x} + \Lambda$ with $V^{-1} = S'$.

THEOREM 3.4. *Suppose Assumptions 3.1-3.3. Let $\mathcal{M}_n$ denote the set of minimizers ζ of $\hat\sigma^2_\zeta$. Then*

(i) $\max_{\zeta\in\mathcal{M}_n} h^{1/2}|\hat\sigma^2_\zeta - \hat\sigma^2_{\zeta_n}| = \mathrm{o_P}(1)$,

(ii) $\max_{\zeta\in\mathcal{M}_n} |(\sum_{i\in\zeta} x_{in}x'_{in})^{1/2}(\hat\beta_\zeta - \beta) - (\sum_{i\in\zeta_n} x_{in}x'_{in})^{1/2}(\hat\beta_{\zeta_n} - \beta)| = \mathrm{o_P}(1)$, *where the square root matrices are defined through joint diagonalization.*

Theorem 3.4 generalizes the asymptotic theory for the location-scale case (Berenguer-Rico et al., 2023). The present assumptions are slightly different and more general. For instance, these allow t_d distributions with their polynomial tails. We note again that the proof of Theorem 3.4 is derived under Assumptions 3.1, 3.2, and 3.3 only, no other structure is imposed.

Using Theorem 3.4, the asymptotic distribution of the LTS estimator can be derived from standard OLS results applied to the infeasible OLS estimator on the "good" observations. For instance, suppose standard OLS assumptions to the "good" observations hold, so that for $i \in \zeta_n$, suppose (x_{in}', ε_i) are i.i.d. with finite fourth moments while $\mathsf{E}(\varepsilon_i|x_{in}) = 0$ and $\mathsf{E}(\varepsilon_i^2|x_{in}) = \sigma^2$. Then, we get that $h^{-1}\sum_{i\in\hat{\zeta}} x_{in}x_{in}' \overset{\mathsf{P}}{\to} \Sigma_x$,

$$\hat{\sigma} \overset{\mathsf{P}}{\to} \sigma \qquad \text{and} \qquad \Big(\sum_{i\in\hat{\zeta}} x_{in}x_{in}'\Big)^{1/2}(\hat{\beta}-\beta)/\hat{\sigma} \overset{\mathsf{D}}{\to} \mathsf{N}(0, I_{\dim x}). \tag{3.3}$$

In the latter result, the square root matrix $S' = (\sum_{i\in\hat{\zeta}} x_{in}x_{in}')^{1/2}$ is computed through a joint diagonalization involving the infeasible matrix $\sum_{i\in\zeta_n} x_{in}x_{in}'$. However, the result remains valid if we replace S' by any matrix R' so that $RR' = \sum_{i\in\hat{\zeta}} x_{in}x_{in}'$. Thus, we can choose R' as the symmetric square root of $\sum_{i\in\hat{\zeta}} x_{in}x_{in}'$ or from a Choleski decomposition of that matrix. It is valid to replace S' with R' since any right square root matrix of $\sum_{i\in\hat{\zeta}} x_{in}x_{in}'$ will be of the form $R' = O'S'$ for an orthogonal matrix $OO' = I_{\dim x}$. Indeed, if R and S have the same square, that is $SS' = RR'$, then pre- and post-multiplication by S^{-1} gives $I_{\dim x} = S^{-1}RR'(S')^{-1}$ so that $R'(S')^{-1} = O'$ is orthogonal, whence $R' = O'S'$. Finally, if $S'(\hat{\beta}-\beta)/\hat{\sigma}$ is asymptotically $\mathsf{N}(0, I_{\dim x})$, so is $O'S'(\hat{\beta}-\beta)/\hat{\sigma}$. That aside, in LTS inference, the estimated asymptotic variance is

$$\widehat{\mathsf{Var}}_{LTS}(\hat{\beta}) = \hat{\sigma}^2\Big(\sum_{i\in\hat{\zeta}} x_{in}x_{in}'\Big)^{-1}. \tag{3.4}$$

This expression avoids the squared consistency factor appearing in $\widehat{\mathsf{Var}}_{SLTS}$ in (2.4).

The "good" errors can be heteroscedastic as long as Assumptions 3.1, 3.2, and 3.3 are satisfied. For instance, suppose that $y_i = \alpha + \beta x_i + \sigma\varepsilon_i$ with univariate x_i and $\varepsilon_i|x_i \sim \mathsf{N}(0, x_i^{\omega})$ where $\omega > 2$ for $i \in \zeta_n$. Suppose, $x_i^{-\omega}$ is i.i.d. gamma with shape and inverse scale of $p/2$. Then, for $i \in \zeta_n$, $\varepsilon_i \sim i.i.d.\mathsf{t}_p$. If $p > 4$, then Assumptions 3.1, 3.2, and 3.3 are satisfied, see Appendix C for details. Theorem 3.4 says that in this case the LTS estimator has the same asymptotic distribution as the OLS estimator on the "good" observations. Since these present heteroscedasticity, valid inference requires Eicker–Huber–White standard errors for LTS in this case. In turn, these will be asymptotically equivalent to corrected standard errors for the infeasible OLS estimator.

4. EXAMPLES OF REGRESSORS

4.1. Assumption to Small Regressors: General Remarks

We start with two general remarks about Assumption 3.1(iii).

Remark 4.1. Recall the frequency of small regressors $F_{nh}(a)$ in (3.1). Assumption 3.1(iii) implies that $\hat{\Sigma}_{\zeta} = h^{-1}\sum_{i\in\zeta} x_{in}x_{in}'$ is positive definite in large samples uniformly in all h-sets ζ (Johansen and Nielsen, 2019). Indeed, for all $\delta \neq 0$,

$$\delta'\hat{\Sigma}_\zeta\delta \geq \min_\zeta h^{-1}\sum_{i\in\zeta}\delta' x_{in}x_{in}'\delta 1_{(|x_{in}'\delta|>a)} \geq a^2 \min_\zeta h^{-1}\sum_{i\in\zeta} 1_{(|x_{in}'\delta|>a)}.$$

Since $h^{-1}\sum_{i\in\zeta} 1_{(|x_{in}'\delta|>a)} = 1 - h^{-1}\sum_{i\in\zeta} 1_{(|x_{in}'\delta|\leq a)} \geq 1 - F_{nh}(a)$, we get

$$\delta'\hat{\Sigma}_\zeta\delta \geq a^2\{1-F_{nh}(a)\} \geq a^2\{1-(\xi+\epsilon)\} > 0,$$

with large probability for large n and for $\epsilon < 1-\xi$ and some $a > 0$.

Remark 4.2. The Assumption to $F_{nh}(a)$ involves a supremum over all h-subsamples. We present two bounds for $F_{nh}(a)$ that avoid the supremum over subsets. These two bounds are useful when checking Assumption 3.1(*iii*) in practice.

The first bound to $F_{nh}(a)$ involves the regressors for all observations

$$F_{nh}(a) \leq (n/h)F_{nn}(a), \tag{4.1}$$

noting that $\sum_{i\in\zeta} 1_{(\cdot)} \leq \sum_{i=1}^n 1_{(\cdot)}$. In particular, Assumption 3.1(*iii*) holds whenever $F_{nn}(a) = \mathrm{o}_{\mathsf{P}}(1)$. See Examples 4.1-4.4 below.

The second bound to $F_{nh}(a)$ only involves the regressors of the "good" observations

$$F_{nh}(a) \leq F_{hh}(a) + (n-h)/h, \tag{4.2}$$

with the convention $F_{hh}(a) = \sup_{\delta:|\delta|=1} h^{-1}\sum_{i\in\zeta_n} 1_{(|x_{in}'\delta|\leq a)}$. This bound follows through $\sum_{i\in\zeta} 1_{(\cdot)} = \sum_{i\in\zeta\cap\zeta_n} 1_{(\cdot)} + \sum_{i\in\zeta\cap\zeta_n^c} 1_{(\cdot)} \leq \sum_{i\in\zeta_n} 1_{(\cdot)} + \sum_{i\in\zeta_n^c} 1$. In particular, if $F_{hh}(a) = \mathrm{o}_{\mathsf{P}}(1)$, then the right hand side of (4.2) has limit $\lambda^{-1}-1$, which is strictly smaller than $2-\lambda^{-1}$ whenever $\lambda > 2/3$. This leaves space for choosing a ξ so that Assumption 3.1(*iii*) is satisfied. This bound is useful to study Assumption 3.1(*iii*) under contaminated regressors, as in the running example above.

4.2. Examples

We analyze regressors with respect to the boundedness Assumption 3.1(*iii*) and the tail behavior condition in Assumption 3.3(*iii*).

Example 4.1. (**Polynomial regressors**). Let $x_{in}' = \{1, (i/n)^q\}$ for $q > 0$ or $0 > q > -1/2$. Use (4.1). Then $F_{nn}(a) = \mathrm{o}_{\mathsf{P}}(1)$ (Johansen and Nielsen, 2019, Ex. 3.2, 3.3) and Assumption 3.1(*iii*) follows. Since x_{in} is bounded, Assumption 3.3(*iiia*) holds.

Example 4.2. (**i.i.d. regressors**). Let $x_{in}' = (1, z_{in})$ where z_{in} is i.i.d. with bounded, continuous density. Use (4.1). Then $F_{nn}(a) = \mathrm{o}_{\mathsf{P}}(1)$ (Johansen and Nielsen, 2019, Thm. 3.3). Assumption 3.3(*iiib*) follows if z_{in} has thinner tails than or the same tails as the "good" errors. For instance, z_{in} for $1 \leq i \leq n$ and ε_i for $i \in \zeta_n$ could be normal.

Example 4.3. (**Stationary regressors**). Let $x_{in}' = (1, z_{in})$ with z_{in} a stationary, normal autoregression. Use (4.1). Then $F_{nn}(a) = \mathrm{o}_{\mathsf{P}}(1)$ (Johansen and Nielsen, 2019, Ex. 3.7). Assumption 3.3(*iiib*) follows if the "good" errors are also normal, since the distribution of the intermediate extreme values for stationary, normal autoregressions is the same as for i.i.d. normal variables (Watts, Rootzén, and Leadbetter, 1982, Thm. 3.3).

Example 4.4. (**Random walk**). Let $x_{in}' = (1, z_{in})$ so $z_{in} = n^{-1/2}\sum_{\ell=1}^{i} \psi_i$ with ψ_i i.i.d. multivariate, zero mean normal. Use (4.1). Then $F_{nn}(a) = \mathrm{o_P}(1)$ (Johansen and Nielsen, 2019, Thm. 3.4) and Assumption 3.1(*iii*) follows. The maximum of a normalized random walk converges in distribution so that Assumption 3.3(*iiia*) is satisfied. The normalized estimator $h^{1/2}(\hat{\beta} - \beta)$ has an asymptotic Dickey–Fuller type distribution. The type depends on how "good" and "outlier" errors alternate (Johansen and Nielsen, 2009).

Example 4.5. (**Binary regressors**). Let $x_{in}' = \{1, 1_{(1 \leq \tau n)}\}$. Use (4.1). Here, $F_{nn}(a) = \max(\tau, 1-\tau)$ for small $a > 0$ (Johansen and Nielsen, 2019, Ex. 3.1). Suppose $\max(\tau, 1-\tau) < 2\lambda - 1$, which is satisfied for instance when $\tau = 1/2$ and $\lambda > 3/4$. The inequality (4.1) then shows that $F_{nh}(a) \leq (n/h)F_{nn}(a) < 2 - 1/\lambda - \epsilon + \mathrm{o}(1)$ for small $\epsilon > 0$. Thus, an $\xi < 2 - 1/\lambda$ can be found so that $F_{nh}(a) \leq \xi$ with large probability. Assumption 3.1(*iii*) follows. The regressor is bounded and Assumption 3.3(*iiia*) follows.

If the "good" regressors satisfy the regularity conditions in Assumptions 3.1–3.3, they can be combined with "outlier" regressors without much structure. In particular, if $F_{hh}(a) \to 0$ as $(a, n) \to (0, \infty)$ then Assumption 3.1(*iii*) is satisfied through the bound (4.2) and it suffices to check that the "outlier" regressors do not drift too fast to satisfy Assumption 3.3(*iii*). See DGPs 4-6 in the running example above.

5. SIMULATIONS

We study the finite sample properties of t-tests for $\beta_0 = \beta_1 = 0$ in the linear model

$$y_i = \beta_0 + \beta_1 z_i + \sigma\varepsilon_i. \tag{5.1}$$

We analyze three statistics and the six data generating processes from Section 2.3. We consider sample sizes $n = 25, 100, 400, 1600, 6400$ with $h/n = \lambda = 0.8$ and use 10^4 repetitions. The code was written in Matlab with LTS estimation done using the `mlts.m` code by Agullo et al. (2008).

Tests. We consider t-statistics $t_{k,s} = \hat{\beta}_{k,s}/\mathsf{se}_{k,s}$, where k and s denote parameter and estimation method, respectively. The t-tests reject for $|t_{k,s}| > q$, where q is the normal 97.5% quantile giving a target level of 5%. We study three estimators so that $s \in \{OLS, LTS, SLTS\}$.

The full sample OLS estimator is $\hat{\beta}_{OLS} = (\sum_{i=1}^{n} x_i x_i')^{-1}(\sum_{i=1}^{n} x_i y_i)$ with $x_i = (1, z_i)'$ while se_{OLS}^2 is the product of $\hat{\sigma}_{OLS}^2 = (n-2)^{-1}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta}_{OLS})^2$ and the relevant diagonal element of $(\sum_{i=1}^{n} x_i x_i')^{-1}$.

The LTS regression estimator $\hat{\beta}$ is given in (2.3) and will be applied with $h/n = 0.8$. We consider two procedures for testing for zero-coefficients using $\hat{\beta}$:

Table 1. Simulated rejection frequencies for nominal 5% tests on intercepts

Method	n	DGP1	DGP2	DGP3	DGP4	DGP5	DGP6
OLS	25	0.060	0.083	0.074	0.218	0.378	0.262
	100	0.053	0.080	0.128	0.732	0.983	0.628
	400	0.050	0.104	0.323	1.000	1.000	0.888
	1600	0.055	0.162	0.741	1.000	1.000	0.948
	6400	0.049	0.293	0.979	1.000	1.000	0.900
LTS	25	0.377	0.268	0.084	0.581	0.063	0.064
	100	0.389	0.177	0.058	0.820	0.053	0.053
	400	0.392	0.113	0.052	0.674	0.050	0.050
	1600	0.400	0.069	0.048	0.263	0.049	0.067
	6400	0.389	0.053	0.050	0.052	0.047	0.918
SLTS	25	0.039	0.017	0.002	0.169	0.000	0.001
	100	0.042	0.003	0.000	0.608	0.000	0.000
	400	0.051	0.000	0.000	0.654	0.000	0.000
	1600	0.050	0.000	0.000	0.220	0.000	0.017
	6400	0.048	0.000	0.000	0.002	0.000	0.724

Note: In all cases, the LTS estimator is computed with $h = 0.8n$ observations. In DGP1, with no contamination, h is chosen wrongly. In all other DGPs, h is chosen correctly.

LTS approach and SLTS approach. First, in the LTS model, we have that $\hat{\beta}$ is asymptotically normal with estimated asymptotic variance given in (3.4). From this we can derive standard errors se_{LTS}, say, and form t-statistics for testing $\beta_0 = 0$ and $\beta_1 = 0$. Second, under the ϵ-contaminated normal model, the SLTS inference procedure is used. The estimated asymptotic variance is given in (2.4) leading to standard errors se_{SLTS} and another set of t-statistics. We note that the standard errors satisfy $\mathsf{se}_{SLTS} = \mathsf{se}_{LTS}/\varsigma^2_{0.8}$, where $\varsigma^2_{0.8} = 0.438$, so that rejection frequencies for SLTS inference are always smaller than for LTS inference.

Data Generating Processes (DGPs). We consider the six DGPs described in Section 2.3, see also Figure 1.

Tables 1 and 2 report simulated rejection frequencies for nominal 5% tests on the intercept and the slope, respectively. Results are based on 10^4 repetitions. The Monte Carlo standard error is 0.2% for correctly sized tests.

DGP 1 has no contamination. Both OLS and SLTS statistics perform well in small samples. The LTS statistic uses the wrong $h = 80$. Hence, it is oversized for all samples sizes. LTS with the correct $h = n$ in DGP 1 is OLS, which has excellent size control. These results are seen both for intercept and slope.

DGPs 2–3 have contamination in the errors, but not in the regressors. The LTS test has empirical size approaching 5% as the sample size increases for both intercept and slope. The LTS procedure works better in finite samples under DGP 3 than DGP 2, since DGP 3 has more separation of "good" and "outlier" errors. The OLS procedure performs differently for intercept and slope. Specifically, the

Table 2. Simulated rejection frequencies for nominal 5% tests on slopes

Method	n	DGP1	DGP2	DGP3	DGP4	DGP5	DGP6
OLS	25	0.064	0.057	0.059	0.754	0.999	1.000
	100	0.051	0.052	0.051	1.000	1.000	1.000
	400	0.053	0.049	0.048	1.000	1.000	1.000
	1600	0.047	0.049	0.049	1.000	1.000	1.000
	6400	0.050	0.051	0.050	1.000	1.000	1.000
LTS	25	0.366	0.290	0.092	0.905	0.065	0.066
	100	0.374	0.191	0.060	0.877	0.050	0.050
	400	0.386	0.135	0.053	0.683	0.052	0.052
	1600	0.390	0.098	0.051	0.279	0.054	0.069
	6400	0.398	0.084	0.053	0.065	0.049	1.000
SLTS	25	0.035	0.023	0.003	0.628	0.000	0.001
	100	0.039	0.003	0.000	0.859	0.000	0.000
	400	0.046	0.000	0.000	0.655	0.000	0.000
	1600	0.046	0.000	0.000	0.220	0.000	0.018
	6400	0.047	0.000	0.000	0.002	0.000	1.000

Note: In all cases, the LTS estimator is computed with $h = 0.8n$ observations. In DGP1, with no contamination, h is chosen wrongly. In all other DGPs, h is chosen correctly.

empirical size for the intercept increases with sample size; whereas the empirical size for the slope statistic is approximately 5% for all sample sizes considered. The SLTS tests have empirical size close to zero for almost all sample sizes considered for both intercept and slope.

DGPs 4–5 have leverage points with positive contamination in both errors and regressors. The LTS test has empirical size approaching 5% as the sample size increases, for both intercept and slope. We note that LTS works better in finite samples under DGP 5 than DGP 4, since DGP 5 has more separation of "good" and "outlier" errors. The near "outliers" configuration in DGP 4 requires larger sample sizes for the asymptotic approximation to come through. The OLS test has empirical size approaching one for both intercept and slope. The SLTS test has a more complicated behavior. For DGP 4, the size first increases for both intercept and slope and then decreases to near zero for large samples. For DGP 5, the size is near zero in all cases.

DGP 6 has positive contamination in the errors and positively contaminated regressors which are growing at the order of $n^{1/2}$ and larger than the largest "good" regressors. The leverage points therefore become relatively closer to the regression line as n grows and are designed to violate Assumptions 3.2(*i*) and 3.3(*iii*) in the LTS asymptotics. The LTS test has empirical size around 5% for most sample sizes, but the size jumps to near unity for the largest sample size. This supports the idea of looking for "outliers" in the regressors before using the LTS estimator (Rousseeuw, 1994). The OLS test has an empirical size that is steadily growing

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

with the sample size for the intercept and constantly at unity for the slope. The SLTS test has empirical size close to 0% for most sample sizes, but the size jumps to near unity for the largest sample size.

6. DISCUSSION: LTS IN PRACTICE

6.1. Inference

The asymptotic theory provided in the previous sections together with the simulation evidence make it clear that inference using LTS depends on the underlying model that generated the data. We have seen that critical values to test for significance of regression parameters vary depending on whether an LTS model or an ϵ-contamination model is the relevant choice. The LTS model will be appropriate for some data sets. For other data sets, ϵ-contamination—the SLTS model—could be attractive despite the nuisance parameters in the inference.

In order to discriminate between models, misspecification tests can be used in practice. The LTS estimator divides the sample into two groups: the "good" and the "outliers". Testing the distributional properties of the "good" LTS residuals can guide users in choosing the relevant model for the data at hand and, hence, in conducting valid inference. For instance, if the "good" errors are normal, then the new LTS inference derived above can be used. Given Theorem 3.4, we suspect that a standard cumulant based test on the "good" residuals will be asymptotically valid to test for normality, but this is yet to be proved. Alternatively, a test for truncated normality after LTS estimation is proposed by Berenguer-Rico and Nielsen (2023). Evidence of truncated normality in this context indicates that an ϵ-tail contamination model where the errors have an i.i.d. structure with a common distribution that is normal in the middle but has flexible tails would be more appropriate, see Berenguer-Rico and Nielsen (2023) for details. In the setting of ϵ-tail contamination, subsequent inference on the regression parameters requires taking into account the presence of nuisance parameters in the asymptotic distribution of the LTS estimator.

In contrast, if there is evidence in favor of an LTS model structure, like for instance, evidence of untruncated normal "good" errors, then Theorem 3.4 tells us that subsequent inference on the regression parameters can be conducted as usual with the LTS estimator. That is, as if we had known which were the "good" observations and we had used OLS on those. If evidence of heteroscedasticity or autocorrelation of the "good" errors is suspected, then Theorem 3.4 indicates that it might be a good idea to account for these features.

6.2. Choosing *h*

When using the LTS estimator in practice, the user has to choose h, the number of "good" observations. Some methodologies to choose h exist in the literature but not so many. We describe some of these as presented in Berenguer-Rico et al. (2023).

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

The traditional approach for choosing h is the *index plot* method of Rousseeuw and Leroy (1987), Rousseeuw and Hubert (1997), which works as follows. First, compute the LTS estimator with $h = n/2$, approximately. Second, standardize all n residuals. This is done using an estimator for the scale that includes a consistency factor computed under the assumption of normal errors and no contamination which leads to the SLTS inference described in Section 2.2. Third, keep observations with absolute scaled residuals smaller than 2.5. This third step determines the value of h. Note that this method will typically declare some observations as "outliers" even when there is no contamination and the errors are all, say, normal.

To improve upon the *index plot* methodology, Berenguer-Rico et al. (2023) investigate two alternative approaches for estimating h based on the LTS model. The first approach uses information criteria in the context of a location-scale model, that is a regression model without regressors. The standard AIC/BIC idea of penalizing the estimated residual variance is not practical, because the resulting estimator of $\lambda = h/n$ appears to converge rather slowly at a $\log\log n$ rate.

The second approach estimates h by choosing the h value that minimizes the cumulant based normality test statistic. Berenguer-Rico et al. (2023) argued that this method delivers a consistent estimation of the proportion of "good" observations, λ, in the location-scale case with a $\log n$ rate.

Consistency of the latter two approaches (information criteria or cumulant based normality test statistic) is argued as follows. Suppose data are generated with $h_\circ$ "good" observations. Consider the three scenarios where: (i) $h = h_\circ$; (ii) $h > h_\circ$; and (iii) $h < h_\circ$. Berenguer-Rico et al. (2023) showed that in the location-scale case, if $h = h_\circ$, then the LTS estimator is $h^{1/2}$-consistent, asymptotically normal, and free of nuisance parameters. It is argued that when $h > h_\circ$, then the LTS estimator in the LTS location-scale model will be divergent and so will be the criterion function for choosing h. Moreover, when $h < h_\circ$, then the LTS estimator truncates the distribution of the residuals and the sample moments converge to a truncated distribution. The criterion functions for estimating h do not account for this truncation. Therefore, the criterion functions diverge in this case too. These considerations lead to a consistency argument in either approach for the location-scale case.

Whether these arguments extend to regression is an open question, but some intuition can be drawn from the above results. On the one hand, if $h > h_\circ$, then "outliers" are included in the set of estimated "good" errors making LTS potentially unbounded, see Remark 3.4 in Section 3.1. On the other hand, if $h < h_\circ$, then the LTS estimator truncates the distribution of the "good" residuals and the theory above does not apply, see the simulation results in Tables 1 and 2 for DGP 1 using LTS where $h = 80 < h_\circ = 100$. Significance tests for both intercept and slope are heavily oversized in that case. Hence, pinning down the correct h is crucial in practice. Extending the h estimators in Berenguer-Rico et al. (2023) to the regression context would make LTS even more useful in practice. The results in this paper are a stepping stone in this direction.

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

APPENDIX

A. Proofs

A.1. Boundedness

Proof of Theorem 3.1. We adapt the proof of Johansen and Nielsen (2019).

(a) *Overview*. We want to prove that $\max_{\zeta\in\mathcal{M}_n}|\hat{\beta}_\zeta-\hat{\beta}_{\zeta_n}|=\mathsf{O}_\mathsf{P}(1)$, where $\mathcal{M}_n$ is the set of minimizers. That is, $\forall\epsilon>0, \exists B_0, n_0>0, \forall n>n_0$ and with $\mathcal{A}_n=(\max_{\zeta\in\mathcal{M}_n}|\hat{\beta}_\zeta-\hat{\beta}_{\zeta_n}|>B_0)$, then $\mathsf{P}(\mathcal{A}_n)<\epsilon$.

Defining the set $\mathcal{A}_{n\zeta}=(|\hat{\beta}_\zeta-\hat{\beta}_{\zeta_n}|>B_0)$, we can write $\mathcal{A}_n=\cup_{\zeta\in\mathcal{M}_n}\mathcal{A}_{n\zeta}$.

Any minimizer $\zeta\in\mathcal{M}_n$ has a residual variance satisfying $\hat{\sigma}^2_\zeta\le\hat{\sigma}^2_{\zeta_n}$. Let $\mathcal{Z}_n$ be the set of all possible ζ and define the set $\mathcal{B}_{n\zeta}=(\hat{\sigma}^2_\zeta\le\hat{\sigma}^2_{\zeta_n})$. Since $\mathcal{B}_{n\zeta}$ contains all minimizers, $\zeta\in\mathcal{M}_n$ and some non-minimizers, we get $\mathcal{A}_n\subset\cup_{\zeta\in\mathcal{Z}_n}(\mathcal{A}_{n\zeta}\cap\mathcal{B}_{n\zeta})$.

Given an $\epsilon>0$, we will find a $B_0>0$ and sets $\mathbb{C}_n$ with probability $\mathsf{P}(\mathbb{C}_n)\ge1-\epsilon$. On $\mathbb{C}_n$, we will argue deterministically that if $|\hat{\beta}_\zeta-\hat{\beta}_{\zeta_n}|>B_0$ for some ζ then $\hat{\sigma}^2_\zeta\ge(1+\epsilon)\hat{\sigma}^2_{\zeta_n}>\hat{\sigma}^2_{\zeta_n}$. Thus, such a ζ cannot be a minimizer. Hence, on $\mathbb{C}_n$, the intersection, $\mathcal{A}_{n\zeta}\cap\mathcal{B}_{n\zeta}$, that is $(\mathcal{A}_{n\zeta}\cap\mathcal{B}_{n\zeta}\cap\mathbb{C}_n)$, is empty. We get

$$(\mathcal{A}_{n\zeta}\cap\mathcal{B}_{n\zeta})=(\mathcal{A}_{n\zeta}\cap\mathcal{B}_{n\zeta}\cap\mathbb{C}_n)\cup(\mathcal{A}_{n\zeta}\cap\mathcal{B}_{n\zeta}\cap\mathbb{C}_n^c)=(\mathcal{A}_{n\zeta}\cap\mathcal{B}_{n\zeta}\cap\mathbb{C}_n^c)\subset\mathbb{C}_n^c.$$

We can then bound $\mathcal{A}_n\subset\cup_{\zeta\in\mathcal{Z}_n}(\mathcal{A}_{n\zeta}\cap\mathcal{B}_{n\zeta})\subset\mathbb{C}_n^c$, so that $\mathsf{P}(\mathcal{A}_n)\le\mathsf{P}(\mathbb{C}_n^c)<\epsilon$.

(b) *Criterion function*. Given a set ζ we find the least squares estimator

$$\hat{\beta}_\zeta=(\sum_{i\in\zeta}x_{in}x_{in}')^{-1}\sum_{i\in\zeta}x_{in}y_i=\beta+(\sum_{i\in\zeta}x_{in}x_{in}')^{-1}\sum_{i\in\zeta}x_{in}\varepsilon_i\sigma$$

using the model equation (2.1). The scaled residuals are $\tilde{\varepsilon}_{\zeta i}=(y_i-x_{in}'\hat{\beta}_\zeta)/\sigma$, so that $h\hat{\sigma}^2_\zeta=\sum_{i\in\zeta}(y_i-x_{in}'\hat{\beta}_\zeta)^2=\sigma^2\sum_{i\in\zeta}\tilde{\varepsilon}^2_{\zeta i}$.

For $\zeta=\zeta_n$ write $\tilde{\varepsilon}_i$ for $\tilde{\varepsilon}_{\zeta_n i}$. For general ζ write $\tilde{\varepsilon}_{\zeta i}=\varepsilon_i-x_{in}'(\hat{\beta}_\zeta-\beta)/\sigma$. Add and subtract $x_{in}'\hat{\beta}_{\zeta_n}/\sigma$ to get $\tilde{\varepsilon}_{\zeta i}=\varepsilon_i-x_{in}'(\hat{\beta}_{\zeta_n}-\beta)/\sigma-x_{in}'(\hat{\beta}_\zeta-\hat{\beta}_{\zeta_n})/\sigma$ and in turn $\tilde{\varepsilon}_{\zeta i}=\tilde{\varepsilon}_i-x_{in}'(\hat{\beta}_\zeta-\hat{\beta}_{\zeta_n})/\sigma$. Introduce polar coordinates with length $\hat{\ell}_\zeta=|\hat{\beta}_\zeta-\hat{\beta}_{\zeta_n}|/\sigma$ and direction $\hat{\delta}_\zeta=(\hat{\beta}_\zeta-\hat{\beta}_{\zeta_n})/|\hat{\beta}_\zeta-\hat{\beta}_{\zeta_n}|$ when $\hat{\ell}_\zeta>0$. When $\hat{\ell}_\zeta=0$ the direction $\hat{\delta}_\zeta$ can be chosen as an arbitrary vector of unit length. Thus, $\hat{\beta}_\zeta-\hat{\beta}_{\zeta_n}=\hat{\ell}_\zeta\hat{\delta}_\zeta\sigma$ and $|\hat{\delta}_\zeta|=1$. The residuals satisfy $\tilde{\varepsilon}_{\zeta i}=\tilde{\varepsilon}_i-\hat{\ell}_\zeta x_{in}'\hat{\delta}_\zeta$, so that

$$h\hat{\sigma}^2_\zeta=\sigma^2\sum_{i\in\zeta}(\tilde{\varepsilon}_i-\hat{\ell}_\zeta x_{in}'\hat{\delta}_\zeta)^2. \tag{A.1}$$

(c) *Bounding residuals under constraints to* ε_i, $x_{in}'\hat{\delta}_\zeta$ *and* $\hat{\ell}_\zeta$. We will later choose $A_0, a_0, C_0>0$. Let $B_{n0}=(A_0+C_0\hat{\sigma}_{\zeta_n}/\sigma)/a_0$. Consider $|\tilde{\varepsilon}_i|\le A_0$ and $|x_{in}'\hat{\delta}_\zeta|>a_0$ and $\hat{\ell}_\zeta>B_{n0}$. Then, by the reverse triangle inequality, $|x-y|\ge|(|x|-|y|)|\ge|y|-|x|$,

$$|\tilde{\varepsilon}_i-\hat{\ell}_\zeta x_{in}'\hat{\delta}_\zeta|\ge\hat{\ell}_\zeta|x_{in}'\hat{\delta}_\zeta|-|\tilde{\varepsilon}_i|>B_{n0}a_0-A_0\ge C_0\hat{\sigma}_{\zeta_n}/\sigma. \tag{A.2}$$

(d) *Bounding residual variance for large* $\hat{\ell}_\zeta$. Apply the expression for $\hat{\sigma}^2_\zeta$ in (A.1). Delete summands of $\hat{\sigma}^2_\zeta$ for which $|\tilde{\varepsilon}_i|>A_0$ or $|x_{in}'\hat{\delta}_\zeta|\le a_0$ and consider only values of ζ with large

$\hat{\ell}_\zeta > B_{n0}$ to get the lower bound

$$h\hat{\sigma}^2_\zeta \geq 1_{(\hat{\ell}_\zeta > B_{n0})}\sigma^2 \sum_{i\in\zeta}(\tilde{\varepsilon}_i - \hat{\ell}_\zeta x'_{in}\hat{\delta}_\zeta)^2 1_{(|\tilde{\varepsilon}_i|\leq A_0)} 1_{(|x'_{in}\hat{\delta}_\zeta|>a_0)}.$$

Now, for $\hat{\ell}_\zeta > B_{n0}$ we can apply (A.2) to get the further bound

$$h\hat{\sigma}^2_\zeta \geq 1_{(\hat{\ell}_\zeta > B_{n0})} C_0^2 \hat{\sigma}^2_{\zeta_n} \sum_{i\in\zeta} 1_{(|\tilde{\varepsilon}_i|\leq A_0)} 1_{(|x'_{in}\hat{\delta}_\zeta|>a_0)}.$$

Use that for sets $\mathbb{A}$ and $\mathbb{B}$ then $1_{\mathbb{A}\cap\mathbb{B}} = 1_{\mathbb{A}} - 1_{\mathbb{A}\cap\mathbb{B}^c} \geq 1_{\mathbb{A}} - 1_{\mathbb{B}^c}$ so that

$$h\hat{\sigma}^2_\zeta \geq 1_{(\hat{\ell}_\zeta > B_{n0})} C_0^2 \hat{\sigma}^2_{\zeta_n} \Big\{ \sum_{i\in\zeta} 1_{(|\tilde{\varepsilon}_i|\leq A_0)} - \sum_{i\in\zeta} 1_{(|x'_{in}\hat{\delta}_\zeta|\leq a_0)} \Big\}. \tag{A.3}$$

For each sum in (A.3), we find bounds not depending on ζ. The first sum satisfies, noting that $1_{(|\tilde{\varepsilon}_i|\leq A_0)} = 1 - 1_{(|\tilde{\varepsilon}_i|>A_0)}$,

$$\sum_{i\in\zeta} 1_{(|\tilde{\varepsilon}_i|\leq A_0)} \geq \sum_{i\in\zeta\cap\zeta_n} 1_{(|\tilde{\varepsilon}_i|\leq A_0)} = \#(\zeta\cap\zeta_n) - \sum_{i\in\zeta\cap\zeta_n} 1_{(|\tilde{\varepsilon}_i|>A_0)}.$$

Note that $\#(\zeta\cap\zeta_n) = \#\zeta - \#(\zeta\cap\zeta_n^c) \geq \#\zeta - \#\zeta_n^c$. Since $\#\zeta = h$ and $\#\zeta_n^c = n-h$, then $\#(\zeta\cap\zeta_n) \geq 2h-n$. Further, by summing over additional non-negative elements, we have that $\sum_{i\in\zeta\cap\zeta_n} 1_{(|\tilde{\varepsilon}_i|>A_0)} \leq \sum_{i\in\zeta_n} 1_{(|\tilde{\varepsilon}_i|>A_0)}$. The inequality $1_{(|\tilde{\varepsilon}_i|>A_0)} \leq \tilde{\varepsilon}_i^2/A_0^2$ gives the further bound $\sum_{i\in\zeta_n}\tilde{\varepsilon}_i^2/A_0^2$. Since $\tilde{\varepsilon}_i$ are the residuals from OLS regression on ζ_n, we get $\sum_{i\in\zeta_n}\tilde{\varepsilon}_i^2 \leq \sum_{i\in\zeta_n}\varepsilon_i^2$. Thus, the first sum in (A.3) satisfies $\sum_{i\in\zeta} 1_{(|\tilde{\varepsilon}_i|\leq A_0)} \geq 2h-n-\sum_{i\in\zeta_n}\varepsilon_i^2/A_0^2$.

For the second sum in (A.3), replace $\hat{\delta}_\zeta$ by an arbitrary δ, take supremum over δ and take maximum over sets ζ of length h to get the bound

$$\sum_{i\in\zeta} 1_{(|x'_{in}\hat{\delta}_\zeta|\leq a_0)} \leq \max_{\zeta:\#\zeta=h} \sup_{|\delta|=1} \sum_{i\in\zeta} 1_{(|x'_{in}\delta|\leq a_0)} = hF_{nh}(a_0).$$

Insert the bounds in (A.3) to get, uniformly in ζ satisfying $\hat{\ell}_\zeta > B_{n0}$, that

$$\hat{\sigma}^2_\zeta \geq 1_{(\hat{\ell}_\zeta > B_{n0})} C_0^2 \hat{\sigma}^2_{\zeta_n} \Big\{ \frac{2h-n}{h} - A_0^{-2}\frac{1}{h}\sum_{i\in\zeta_n}\varepsilon_i^2 - F_{nh}(a_0) \Big\}. \tag{A.4}$$

(e) *Probability argument*. We construct sets $\mathbb{C}_n$ with large probability.

Assumption 3.1(*i*) has $h/n \to \lambda > 1/2$. Thus, $(2h-n)/h \to 2-\lambda^{-1} > 0$.

Assumption 3.1(*ii*) states that $h^{-1}\sum_{i\in\zeta_n}\varepsilon_i^2 = \mathrm{O_P}(1)$. This implies that $\hat{\sigma}^2_{\zeta_n}/\sigma^2 = \mathrm{O_P}(1)$, see Remark 3.1.

Assumption 3.1(*iii*) states that $\lim_{(a,n)\to(0,\infty)} \mathsf{P}\{F_{nh}(a) > \xi\} = 0$ for some $\xi < 2-\lambda^{-1}$.

These assumptions show that for all $\epsilon > 0$ there exists $a_0, A_0, n_0 > 0$ and sets $\mathbb{C}_n$ with $\mathsf{P}(\mathbb{C}_n) \geq 1-\epsilon$ for all $n > n_0$ so that on $\mathbb{C}_n$ we have

$$\frac{1}{h}\sum_{i\in\zeta_n}\varepsilon_i^2 \leq \epsilon A_0^2 \qquad \text{and} \qquad \hat{\sigma}^2_{\zeta_n}/\sigma^2 \leq A_0 \qquad \text{and} \qquad F_{nh}(a_0) \leq \xi.$$

Now, choose $C_0^2 = (1+\epsilon)/(2-\lambda^{-1}-2\epsilon-\xi)$, noting that $C_0^2 > 0$ for small ϵ since $\xi < 2-\lambda^{-1}$. Let $B_0 = (A_0 + C_0A_0)/a_0$ so that $B_0 \geq B_{n0}$ on $\mathbb{C}_n$.

(f) *Bound residual variance on* $\mathbb{C}_n$. As argued in (*a*), consider any ζ with $\hat{\ell}_\zeta = |\hat{\beta}_\zeta - \hat{\beta}_{\zeta_n}| > B_0 \geq B_{n0}$. Apply the constraints defining $\mathbb{C}_n$ to the lower bound for $\hat{\sigma}^2_\zeta$ in (A.4) to get the bound

$$\hat{\sigma}^2_\zeta \geq C_0^2 \hat{\sigma}^2_{\zeta_n}\{(2-\lambda^{-1}-\epsilon)-\epsilon-\xi\} = (1+\epsilon)\hat{\sigma}^2_{\zeta_n}$$

on the set $\mathbb{C}_n$. Thus, this ζ cannot be a minimizer since minimizers satisfy $\hat{\sigma}^2_\zeta \leq \hat{\sigma}^2_{\zeta_n}$. This is what had to be proved as outlined in item (*a*). □

A.2. Consistent Selection of "Good" Observations

For a matrix m let $\|m\|$ be the spectral norm. Thus, $\|m\|^2 = \max \text{eigen}(m'm)$. If the matrices m_1, m_2 are conformable then $\|m_1 m_2\| \leq \|m_1\|\|m_2\|$.

Proof of Theorem 3.2. We note that for any minimizer, $\zeta \in \mathcal{M}_n$, then $\hat{\sigma}^2_\zeta \leq \hat{\sigma}^2_{\zeta_n}$.

We construct a high probability set $\mathbb{D}_n$, where we can deterministically bound certain statistics. Assumptions 3.1, 3.2(*ii*) along with Remark 3.1 and Theorem 3.1 show that $\max_{\zeta \in \mathcal{M}_n} |\hat{\beta}_\zeta|$ and $\hat{\sigma}^2_{\zeta_n}$ are $\mathrm{O_P}(1)$. Assumption 3.2(*i*) is that $\|\sum_{i=1}^n x_{in}x'_{in}\| = \mathrm{O_P}(n) = \mathrm{O_P}(h)$. Thus, for all $\epsilon > 0$ there exist $C, n_0 > 0$ and a sequence of sets $\mathbb{D}_n$ with $\mathsf{P}(\mathbb{D}_n) > 1-\epsilon$ for all $n > n_0$, so that on $\mathbb{D}_n$

$$\max_{\zeta \in \mathcal{M}_n} |\hat{\beta}_\zeta - \beta|/\sigma \leq C, \quad \hat{\sigma}^2_{\zeta_n}/\sigma^2 \leq C, \quad \|\sum_{i=1}^n x_{in}x'_{in}\| \leq Ch. \tag{A.5}$$

For a minimizer ζ, we expand the least squares residual variance as

$$h\hat{\sigma}^2_\zeta = \sigma^2 \sum_{i\in\zeta} \varepsilon_i^2 - (\hat{\beta}_\zeta - \beta)'(\sum_{i\in\zeta} x_{in}x'_{in})(\hat{\beta}_\zeta - \beta). \tag{A.6}$$

The first term satisfies $\sum_{i\in\zeta}\varepsilon_i^2 \geq \sum_{i\in\zeta\cap\zeta_n^c}\varepsilon_i^2 \geq \#(\zeta\cap\zeta_n^c)\min_{j\notin\zeta_n}\varepsilon_j^2$. For the second term, $\|\sum_{i\in\zeta} x_{in}x'_{in}\| \leq \|\sum_{i=1}^n x_{in}x'_{in}\| \leq Ch$ and $|\hat{\beta}_\zeta - \beta| \leq C\sigma$. Further, $C\sigma^2 \geq \hat{\sigma}^2_{\zeta_n} \geq \hat{\sigma}^2_\zeta$. Thus, on the set $\mathbb{D}_n$, we get

$$hC \geq h\hat{\sigma}^2_\zeta/\sigma^2 \geq \#(\zeta\cap\zeta_n^c)\min_{j\notin\zeta_n}\varepsilon_j^2 - C^3 h. \tag{A.7}$$

On $\mathbb{D}_n$, solve to get $\#(\zeta\cap\zeta_n^c) \leq (C+C^3)h/\min_{j\notin\zeta_n}\varepsilon_j^2$, uniformly in $\zeta \in \mathcal{M}_n$. □

A.3. Improving the Rate of Consistency

When improving the consistency rate we will bound terms like $\sum_{i\in\zeta\cap\zeta_n^c} x_{in}x'_{in}$. The bounds should be uniform in ζ where the number of misclassifications, $\#(\zeta\cap\zeta_n^c)$, is bounded by some sequence g_n. Thus, everywhere, $g_n > 0$ is a sequence in n not depending on ζ. When applying the bounds we will first choose $g_n = Ch/m_n^2$ and later $g_n = Ch^\theta$. The bounds will be expressed in terms of

$$\mathcal{R}_{g_n} = \sum_{i>n-g_n}^n x_{(i)}^2 \qquad \text{and} \qquad \mathcal{S}_{g_n} = \sum_{i>n-g_n}^n \phi_i^2,$$

where $x_{(i)}$ and ϕ_i are the increasing order statistics of $|x_{in}|$ and $x'_{in}(\sum_{\ell\in\zeta_n} x_{\ell n}x'_{\ell n})^{-1}x_{in}$.

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

LEMMA A.1. (a) *Suppose Assumption 3.3 parts* $(ia, iiia)$ *or parts* $(ia, id, iiib)$. *Then,* $\forall C > 0$: $\mathcal{R}_{g_n} = \mathrm{o}_{\mathsf{P}}(h)$ *for* $g_n \leq Ch/m_n^2$.
(b) *Suppose Assumption 3.3*(ia, id, iii). *Then,* $\forall C, \theta > 0$: $\mathcal{R}_{g_n} = \mathrm{O}_{\mathsf{P}}(n^{\eta+\theta})L(n)$ *for* $g_n \leq Ch^{\theta}$.
(c) *Suppose Assumption 3.1*(iii). *Then* $\mathcal{S}_{g_n} = \mathcal{R}_{g_n}\mathrm{O}_{\mathsf{P}}(n^{-1})$.

Remark A.1. For Lemma A.1, and hence for Theorem 3.3, it suffices that $0 < \eta < 1$ in Assumption 3.3(id, iii).

Proof of Lemma A.1. (a) with Assumption 3.3$(iiia)$, where $x_{(n)}^2 = \mathrm{O}_{\mathsf{P}}(1)$. Assumption 3.3$(ia)$ has $m_n^2 \to \infty$ so that $x_{(n)}^2 = \mathrm{o}_{\mathsf{P}}(m_n^2)$. Thus, $\mathcal{R}_{g_n}/g_n \leq x_{(n)}^2 = \mathrm{o}_{\mathsf{P}}(m_n^2)$.

(a) with Assumption 3.3$(iiib)$ where $x_{(n)}^2 = \mathrm{O}_{\mathsf{P}}(m_n^2)$. Since $\mathcal{R}_{g_n}$ is increasing in g_n, it suffices to consider $g_n = Ch/m_n^2$. Assumption 3.3(ia, id) have $m_n^2 \to \infty$ but $m_n^2 = \mathrm{O}_{\mathsf{P}}(n^{\eta})L(n)$. Further, by Assumption 3.3$(iiib)$, then for all $0 < \delta < 1$ exists $r < 1 - \eta$ so that $x_{(n-\lfloor n^r \rfloor)}^2 / x_{(n)}^2 \leq \delta\{1 + \mathrm{o}_{\mathsf{P}}(1)\}$. For such r, then $n^r \leq \delta n^{1-\eta}/L(n)$ for any $\delta > 0$ and large n, since $n^{1-\eta-r}$ dominates the slowly varying function $L(n)$ (Karamata, 1930, p. 45). Now, partition

$$\mathcal{R}_{g_n} = \sum_{i>n-g_n}^{n} x_{(i)}^2 = \sum_{i>n-g_n}^{i \leq n-n^r} x_{(i)}^2 + \sum_{i>n-n^r}^{n} x_{(i)}^2 \leq g_n x_{(n-\lfloor n^r \rfloor)}^2 + n^r x_{(n)}^2. \tag{A.8}$$

Insert first the bounds $x_{(n-\lfloor n^r \rfloor)}^2 \leq \delta x_{(n)}^2\{1 + \mathrm{o}_{\mathsf{P}}(1)\}$ and $n^r \leq \delta n^{1-\eta}/L(n)$ to bound

$$\mathcal{R}_{g_n} \leq g_n \delta x_{(n)}^2\{1 + \mathrm{o}_{\mathsf{P}}(1)\} + x_{(n)}^2 \delta n^{1-\eta}/L(n) = \delta x_{(n)}^2 \big[g_n\{1 + \mathrm{o}_{\mathsf{P}}(1)\} + n^{1-\eta}/L(n)\big].$$

Apply the assumed bound $x_{(n)}^2 = \mathrm{O}_{\mathsf{P}}(m_n^2)$. In the square bracket term, insert $g_n = Ch/m_n^2$ for the first summand so that the m_n^2 terms cancel and use the bound $m_n^2 = \mathrm{O}_{\mathsf{P}}(n^{\eta})L(n)$ for the second term to get

$$\mathcal{R}_{g_n} \leq \mathrm{O}_{\mathsf{P}}(\delta)\big[Ch\{1 + \mathrm{o}_{\mathsf{P}}(1)\} + \mathrm{O}_{\mathsf{P}}(n)\big] = \mathrm{O}_{\mathsf{P}}(\delta)h,$$

since $n \leq 2h$, so that we can take common factor h and simplify the remainder terms. Since $\delta > 0$ can be chosen small, we find $\mathcal{R}_{g_n} = \mathrm{o}_{\mathsf{P}}(h)$.

(b) First bound $\mathcal{R}_{g_n} = \sum_{i>n-g_n}^{n} x_{(i)}^2 \leq g_n x_{(n)}^2$. It suffices to consider $g_n = Ch^{\theta}$. Hence, $\mathcal{R}_{g_n} \leq Ch^{\theta} x_{(n)}^2$. By Assumption 3.3$(ia, iii)$, $x_{(n)}^2 = \mathrm{O}_{\mathsf{P}}(m_n^2)$. Therefore, $\mathcal{R}_{g_n} = h^{\theta}\mathrm{O}_{\mathsf{P}}(m_n^2)$. By Assumption 3.3$(id)$, $m_n^2 = \mathrm{O}_{\mathsf{P}}(n^{\eta})L(n)$. Hence, $\mathcal{R}_{g_n} = h^{\theta}\mathrm{O}_{\mathsf{P}}(n^{\eta})L(n)$. (c) We bound

$$x_{in}'\Big(\sum_{\ell \in \zeta_n} x_{\ell n} x_{\ell n}'\Big)^{-1} x_{in} \leq |x_{in}|^2 \Big\|\Big(\sum_{\ell \in \zeta_n} x_{\ell n} x_{\ell n}'\Big)^{-1}\Big\|.$$

Taking sum over the largest g_n order statistics on the left hand side is less than the sum of the largest g_n order statistics on the right hand side. Thus, we get

$$\mathcal{S}_{g_n} \leq \mathcal{R}_{g_n} \Big\|\Big(\sum_{\ell \in \zeta_n} x_{\ell n} x_{\ell n}'\Big)^{-1}\Big\|. \tag{A.9}$$

The inverse square sum is $\mathrm{O}_{\mathsf{P}}(h^{-1})$ under Assumption 3.1(iii), see Remark 4.1. We then get that $\mathcal{S}_{g_n} = \mathcal{R}_{g_n}\mathrm{O}_{\mathsf{P}}(h^{-1})$. □

The following notation is convenient. Apply the joint diagonalization $\sum_{i\in\zeta_n} x_{in}x_{in}' = SS'$ and $\sum_{i\in\zeta} x_{in}x_{in}' = S(I_{\dim x}+\Lambda)S'$ as discussed in Section 3.3. Define the asymmetric right square roots of $\sum_{i\in\zeta_n} x_{in}x_{in}'$ and its inverse $(\sum_{i\in\zeta_n} x_{in}x_{in}')^{-1} = (S')^{-1}S^{-1}$ as

$$(\sum_{i\in\zeta_n} x_{in}x_{in}')^{1/2} = S', \qquad (\sum_{i\in\zeta_n} x_{in}x_{in}')^{-1/2} = S^{-1}. \tag{A.10}$$

In general, the latter is not the inverse of the former. In a similar fashion, define

$$(\sum_{i\in\zeta} x_{in}x_{in}')^{1/2} = (I_{\dim x}+\Lambda)^{1/2}S', \qquad (\sum_{i\in\zeta} x_{in}x_{in}')^{-1/2} = (I_{\dim x}+\Lambda)^{-1/2}S^{-1}. \tag{A.11}$$

Now, let

$$\begin{aligned}
z_{jn} &= (\sum_{i\in\zeta_n} x_{in}x_{in}')^{-1/2}x_{jn},\\
A_\zeta &= (\sum_{i\in\zeta} x_{in}x_{in}')^{1/2}(\hat\beta_\zeta-\beta)/\sigma = (\sum_{i\in\zeta} x_{in}x_{in}')^{-1/2}\sum_{i\in\zeta} x_{in}\varepsilon_i,\\
B_\zeta &= \sum_{i\in\zeta} z_{in}\varepsilon_i - \sum_{i\in\zeta_n} z_{in}\varepsilon_i,\\
C_\zeta &= (\sum_{i\in\zeta} x_{in}x_{in}')^{-1/2}\{(\sum_{i\in\zeta_n} x_{in}x_{in}')^{1/2}\}' - I_{\dim x},
\end{aligned} \tag{A.12}$$

so that $A_{\zeta_n} = \sum_{i\in\zeta_n} z_{in}\varepsilon_i$.

LEMMA A.2. *The squared difference $|A_\zeta - A_{\zeta_n}|^2$ can be bounded as follows*

$$\frac{1}{3}|A_\zeta - A_{\zeta_n}|^2 \le |B_\zeta|^2(1+\|C_\zeta\|^2) + \|C_\zeta\|^2|\sum_{i\in\zeta_n} z_{in}\varepsilon_i|^2. \tag{A.13}$$

Proof of Lemma A.2. By definition

$$A_\zeta - A_{\zeta_n} = (\sum_{i\in\zeta} x_{in}x_{in}')^{-1/2}\{(\sum_{i\in\zeta_n} x_{in}x_{in}')^{1/2}\}'(\sum_{i\in\zeta} z_{in}\varepsilon_i) - (\sum_{i\in\zeta_n} z_{in}\varepsilon_i).$$

Rewrite as $A_\zeta - A_{\zeta_n} = B_\zeta + C_\zeta B_\zeta + C_\zeta(\sum_{i\in\zeta_n} z_{in}\varepsilon_i)$. The triangle and Jensen's inequalities and the spectral norm sub-multiplicativity give the desired result. □

LEMMA A.3. *Let $\#(\zeta\cap\zeta_n^c)\le g_n$. Then $\sum_{i\in\zeta\cap\zeta_n^c} z_{in}'z_{in}$, $\sum_{i\in\zeta^c\cap\zeta_n} z_{in}'z_{in}$ are at most $\mathcal{S}_{g_n}$.*

Proof of Lemma A.3. By definition $z_{in}'z_{in} = x_{in}'(\sum_{\ell\in\zeta_n} x_{\ell n}x_{\ell n}')^{-1}x_{in}$. As remarked in Section 3.2, we have $\#(\zeta\cap\zeta_n^c) = \#(\zeta^c\cap\zeta_n)$. Since ϕ_i are the increasing order statistics of $z_{in}'z_{in}$ and $\#(\zeta\cap\zeta_n^c)\le g_n$ both sums are bounded by $\mathcal{S}_{g_n}$. □

LEMMA A.4. *Let $\#(\zeta \cap \zeta_n^c) \leq g_n$. The term $|B_\zeta|^2$ can be bounded by*

$$|B_\zeta|^2 \leq 2\Big(\sum_{i\in\zeta\cap\zeta_n^c} \varepsilon_i^2 + \sum_{i\in\zeta^c\cap\zeta_n} \varepsilon_i^2\Big)\mathcal{S}_{g_n}.$$

Proof of Lemma A.4. Decompose B_ζ as

$$B_\zeta = \sum_{i\in\zeta} z_{in}\varepsilon_i - \sum_{i\in\zeta_n} z_{in}\varepsilon_i = \sum_{i\in\zeta\cap\zeta_n^c} z_{in}\varepsilon_i - \sum_{i\in\zeta^c\cap\zeta_n} z_{in}\varepsilon_i,$$

where the second equality follows from cancelling elements with index in $\zeta \cap \zeta_n$. Apply the triangle, Jensen and Cauchy–Schwarz inequalities to get

$$\begin{aligned}|B_\zeta|^2 &\leq \Big(\sum_{i\in\zeta\cap\zeta_n^c} |z_{in}\varepsilon_i| + \sum_{i\in\zeta^c\cap\zeta_n} |z_{in}\varepsilon_i|\Big)^2\\ &\leq 2\Big\{\Big(\sum_{i\in\zeta\cap\zeta_n^c} |z_{in}\varepsilon_i|\Big)^2 + \Big(\sum_{i\in\zeta^c\cap\zeta_n} |z_{in}\varepsilon_i|\Big)^2\Big\}\\ &\leq 2\Big(\sum_{i\in\zeta\cap\zeta_n^c} |z_{in}|^2 \sum_{i\in\zeta\cap\zeta_n^c} \varepsilon_i^2 + \sum_{i\in\zeta^c\cap\zeta_n} |z_{in}|^2 \sum_{i\in\zeta^c\cap\zeta_n} \varepsilon_i^2\Big).\end{aligned}$$

Lemma A.3 bounds $\sum_{i\in\zeta\cap\zeta_n^c} |z_{in}|^2$ and $\sum_{i\in\zeta^c\cap\zeta_n} |z_{in}|^2$ by $\mathcal{S}_{g_n}$. □

LEMMA A.5. *Let $M_\zeta = (\sum_{i\in\zeta_n} x_{in}x_{in}')^{-1/2}\sum_{i\in\zeta} x_{in}x_{in}'\{(\sum_{i\in\zeta_n} x_{in}x_{in}')^{-1/2}\}'$. Suppose $\#(\zeta \cap \zeta_n^c) \leq g_n$. Then, $\|M_\zeta - I_{\dim x}\| \leq 2\mathcal{S}_{g_n}$.*

Proof of Lemma A.5. Since $z_{jn} = (\sum_{i\in\zeta_n} x_{in}x_{in}')^{-1/2}x_{jn}$, we get $M_\zeta = \sum_{i\in\zeta} z_{in}z_{in}'$. Write

$$M_\zeta = \sum_{i\in\zeta_n} z_{in}z_{in}' + (\sum_{i\in\zeta} z_{in}z_{in}' - \sum_{i\in\zeta_n} z_{in}z_{in}'),$$

and note that the first sum satisfies $\sum_{i\in\zeta_n} z_{in}z_{in}' = I_{\dim x}$ while, in the last two sums, we can cancel elements with index in $\zeta \cap \zeta_n$. Hence,

$$M_\zeta = I_{\dim x} + \sum_{i\in\zeta\cap\zeta_n^c} z_{in}z_{in}' - \sum_{i\in\zeta^c\cap\zeta_n} z_{in}z_{in}'. \tag{A.14}$$

Use the spectral norm and the triangle inequality to get that

$$\|M_\zeta - I_{\dim x}\| \leq \sum_{i\in\zeta\cap\zeta_n^c} \|z_{in}z_{in}'\| + \sum_{i\in\zeta^c\cap\zeta_n} \|z_{in}z_{in}'\| = \sum_{i\in\zeta\cap\zeta_n^c} z_{in}'z_{in} + \sum_{i\in\zeta^c\cap\zeta_n} z_{in}'z_{in}.$$

By Lemma A.3, each of the sums is bounded by $\mathcal{S}_{g_n}$. The desired bound follows. □

LEMMA A.6. *Suppose $\#(\zeta \cap \zeta_n^c) \leq g_n$ and $\mathcal{S}_{g_n} \leq 1/4$. Then $\|C_\zeta\| \leq 4\mathcal{S}_{g_n}$.*

Proof of Lemma A.6. Let $M_\zeta = (\sum_{i\in\zeta_n} x_{in}x_{in}')^{-1/2}\sum_{i\in\zeta} x_{in}x_{in}'\{(\sum_{i\in\zeta_n} x_{in}x_{in}')^{-1/2}\}'$ as before and recall that $C_\zeta = (\sum_{i\in\zeta} x_{in}x_{in}')^{-1/2}\{(\sum_{i\in\zeta_n} x_{in}x_{in}')^{1/2}\}' - I_{\dim x}$.

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

Apply the definitions of the square roots stated in (A.10), (A.11), to see that

$$M_\zeta - I_{\dim x} = S^{-1}\{S(I_{\dim x}+\Lambda)S'\}(S')^{-1} - I_{\dim x} = \Lambda,$$
$$C_\zeta = (I_{\dim x}+\Lambda)^{-1/2}S^{-1}S - I_{\dim x} = (I_{\dim x}+\Lambda)^{-1/2} - I_{\dim x}.$$

We will exploit that both matrices are diagonal. Further, Lemma A.5 shows that $\|\Lambda\| = \|M_\zeta - I_{\dim x}\| \le 2\mathcal{S}_{g_n}$. Due to the condition $\mathcal{S}_{g_n} \le 1/4$ we have that the diagonal elements of Λ satisfy $|\lambda_i| \le 1/2$. Due to the diagonality, it suffices to check the bound for $\|C_\zeta\|$ by bounding each diagonal element using the scalar inequality $|(1+\lambda)^{-1/2} - 1| \le 2|\lambda|$ for any scalar $|\lambda| \le 1/2$.

The scalar inequality is equivalent to $1 - 2|\lambda| \le (1+\lambda)^{-1/2} \le 1 + 2|\lambda|$. The upper inequality, for instance, holds by inspection for $\lambda \ge 0$, while, for $\lambda < 0$, we can square the inequality to get $(1+\lambda)^{-1} \le (1-2\lambda)^2$ or equivalently $1 \le (1-2\lambda)^2(1+\lambda) = 1 - 3\lambda + 4\lambda^3$ or equivalently $0 \le (-\lambda)(3-4\lambda^2)$, which is indeed true since $-\lambda > 0$ and $3 - 4\lambda^2 \ge 2$ for $-1/2 \le \lambda < 0$. The lower inequality follows by an analogous argument. □

LEMMA A.7. *Suppose Assumption 3.3(iv). Let* $\mathcal{S}_{g_n} = \mathrm{o_P}(1)$. *Consider all* ζ *so that* $\#(\zeta \cap \zeta_n^c) \le g_n$. *Then,*

$$|A_\zeta - A_{\zeta_n}|^2 \le \Big(\sum_{i\in\zeta\cap\zeta_n^c} \varepsilon_i^2 + \sum_{i\in\zeta^c\cap\zeta_n} \varepsilon_i^2\Big)\mathrm{O_P}(\mathcal{S}_{g_n}) + \mathrm{o_P}(1),$$

where the remainder terms are uniform in ζ .

Proof of Lemma A.7. By Lemma A.2,

$$|A_\zeta - A_{\zeta_n}|^2 \le 3|B_\zeta|^2\{1+\|C_\zeta\|^2\} + 3\|C_\zeta\|^2|\sum_{i\in\zeta_n} z_{in}\varepsilon_i|^2. \tag{A.15}$$

By Assumption 3.3(*iv*), $|\sum_{i\in\zeta_n} z_{in}\varepsilon_i|^2 = (\sum_{i\in\zeta_n} \varepsilon_i z_{in}')(\sum_{i\in\zeta_n} z_{in}\varepsilon_i) = \mathrm{O_P}(1)$. By Lemma A.6 using the Assumption that $\mathcal{S}_{g_n} = \mathrm{o_P}(1)$, we get $\|C_\zeta\| = \mathrm{O_P}(\mathcal{S}_{g_n}) = \mathrm{o_P}(1)$ uniformly in ζ. By Lemma A.4, $|B_\zeta|^2 \le 2(\sum_{i\in\zeta\cap\zeta_n^c} \varepsilon_i^2 + \sum_{i\in\zeta^c\cap\zeta_n} \varepsilon_i^2)\mathcal{S}_{g_n}$. Insert these results into (A.15). □

LEMMA A.8. *Suppose Assumption 3.3(iv). Let* $\mathcal{S}_{g_n} = \mathrm{o_P}(1)$. *Consider all* ζ *so that* $\#(\zeta \cap \zeta_n^c) \le g_n$. *Then,*

$$h(\hat\sigma_\zeta^2 - \hat\sigma_{\zeta_n}^2)/\sigma^2 \ge \{1+\mathrm{o_P}(1)\}\sum_{i\in\zeta\cap\zeta_n^c} \varepsilon_i^2 - \{1+\mathrm{o_P}(1)\}\sum_{i\in\zeta^c\cap\zeta_n} \varepsilon_i^2 + \mathrm{O_P}(1),$$

where all remainder terms are uniform in ζ.

Remark A.2. Note that Lemma A.8 only uses Assumption 3.3(iv). In particular, it is not used that ε_j diverge for $j \notin \zeta_n$.

Proof of Lemma A.8. Write

$$Q_\zeta \;= h(\hat\sigma_\zeta^2 - \hat\sigma_{\zeta_n}^2)/\sigma^2 = \sum_{i\in\zeta} \varepsilon_i^2 - A_\zeta' A_\zeta - \sum_{i\in\zeta_n} \varepsilon_i^2 + A_{\zeta_n}' A_{\zeta_n}. \tag{A.16}$$

Cancelling elements with index in $\zeta \cap \zeta_n$ and note $A'_{\zeta_n} A_{\zeta_n} \geq 0$ to bound

$$Q_\zeta \geq \sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 - \sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 - A'_\zeta A_\zeta. \tag{A.17}$$

Write $A_\zeta = A_{\zeta_n} + (A_\zeta - A_{\zeta_n})$ to get

$$A'_\zeta A_\zeta \leq 2\{A'_{\zeta_n} A_{\zeta_n} + (A_\zeta - A_{\zeta_n})'(A_\zeta - A_{\zeta_n})\} = 2A'_{\zeta_n} A_{\zeta_n} + 2|A_\zeta - A_{\zeta_n}|^2. \tag{A.18}$$

Assumption 3.3(iv) has $A'_{\zeta_n} A_{\zeta_n} = \mathrm{O_P}(1)$. Lemma A.7 using Assumption 3.3(iv) and $\mathcal{S}_{g_n} = \mathrm{o_P}(1)$ uniformly in ζ bounds $|A_\zeta - A_{\zeta_n}|^2 \leq (\sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 + \sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2)\mathrm{o_P}(1) + \mathrm{o_P}(1)$, where the remainders are uniform in ζ. Therefore, the bound (A.18) becomes

$$A'_\zeta A_\zeta \leq \mathrm{O_P}(1) + \Big(\sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 + \sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 \Big) \mathrm{o_P}(1) + \mathrm{o_P}(1). \tag{A.19}$$

Insert (A.19) in (A.17) to get the desired result. □

Lemma A.9. *Suppose Assumptions 3.1(iii), 3.3. Then,* $\forall C > 0$, $0 < \theta < 1 - \eta$, *we have that* $\min_{\zeta : h^\theta \leq \#(\zeta \cap \zeta_n^c) \leq hC/m_n^2} h^{1-\theta}(\hat\sigma_\zeta^2 - \hat\sigma_{\zeta_n}^2) \to \infty$ *in probability.*

Proof of Lemma A.9. Let # be shorthand for $\#(\zeta^c \cap \zeta_n) = \#(\zeta \cap \zeta_n^c)$.

We consider $h^\theta \leq \# \leq g_n$ where $g_n = hC/m_n^2$. We have that $\mathcal{S}_{g_n} = \mathrm{o_P}(1)$, by Lemma A.1($a, c$) using Assumptions 3.1($iii$) and 3.3($ia, id, iii$). Thus, Lemma A.8 using Assumption 3.3(iv) and $\mathcal{S}_{g_n} = \mathrm{o_P}(1)$, shows

$$h(\hat\sigma_\zeta^2 - \hat\sigma_{\zeta_n}^2)/\sigma^2 \geq \{1 + \mathrm{o_P}(1)\} \sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 - \{1 + \mathrm{o_P}(1)\} \sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 + \mathrm{O_P}(1), \tag{A.20}$$

where all remainder terms are uniform in ζ. We show that the lower bound diverges.

The first sum in (A.20) relates to "outliers", which satisfy $\varepsilon_j^2 \geq m_n^2\{1 + \mathrm{o_P}(1)\}$ for $j \notin \zeta_n$ by Assumption 3.3(ii). Thus, $\sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 \geq m_n^2 \#\{1 + \mathrm{o_P}(1)\}$.

The second sum in (A.20) relates to "good" errors. Let $\psi_1 \leq \cdots \leq \psi_h$ be the order statistics of ε_i^2 for $i \in \zeta_n$. Given $\theta > 0$ choose $0 < \rho < \theta$. Since $\lfloor h^\rho \rfloor < \#$, then

$$\sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 \leq \sum_{i = h+1-\#}^{h} \psi_i = \sum_{i = h+1-\#}^{h - \lfloor h^\rho \rfloor} \psi_i + \sum_{i = h - \lfloor h^\rho \rfloor + 1}^{h} \psi_i \leq \# \psi_{h - \lfloor h^\rho \rfloor} + h^\rho \psi_h.$$

For the first term, Assumption 3.3(ic) shows a $C_\rho < 1$ exists so that $\psi_{h - \lfloor h^\rho \rfloor}/m_n^2 \leq C_\rho + \mathrm{o_P}(1)$. Thus, the first term is bounded by $m_n^2 \#\{C_\rho + \mathrm{o_P}(1)\}$.

For the second term, we have $\rho < \theta$ so that $h^\rho = \mathrm{o}(h^\theta)$ while $h^\theta \leq \#$ by construction. Further, Assumption 3.3(ib) shows $\psi_h / m_n^2 = \mathrm{O_P}(1)$. Thus, the second term is bounded by $m_n^2 \# \mathrm{o_P}(1)$. Overall, we get $\sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 \leq m_n^2 \#\{C_\rho + \mathrm{o_P}(1)\}$.

Inserting the above bounds in (A.21), we find

$$h(\hat\sigma_\zeta^2 - \hat\sigma_{\zeta_n}^2)/\sigma^2 \geq m_n^2 \#(1 - C_\rho)\{1 + \mathrm{o_P}(1)\}.$$

Since m_n^2 diverges due to Assumption 3.3(ia), $\# \geq h^\theta$ and $C_\rho < 1$, then $h^{1-\theta}(\hat\sigma_\zeta^2 - \hat\sigma_{\zeta_n}^2) \to \infty$ in probability. □

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

Proof of Theorem 3.3. First, Theorem 3.2 using Assumptions 3.1 and 3.2, shows that $\max_{\zeta\in\mathcal{M}_n}\#(\zeta\cap\zeta_n^c) = \mathrm{O}_{\mathsf{P}}(h/\min_{j\notin\zeta_n}\varepsilon_j^2)$. Here, $\min_{j\notin\zeta_n}\varepsilon_j^2 \geq m_n^2\{1+\mathrm{o}_{\mathsf{P}}(1)\}$ by Assumption 3.3(*ii*). Hence, $\max_{\zeta\in\mathcal{M}_n}\#(\zeta\cap\zeta_n^c) = \mathrm{O}_{\mathsf{P}}(h/m_n^2)$.

Second, Lemma A.9 using Assumptions 3.1(*iii*), 3.3 considers estimators $\hat{\sigma}_\zeta^2$ for index sets ζ that contain a positive number of "outliers" in the range $h^\theta \leq \#(\zeta\cap\zeta_n^c) \leq Ch/m_n^2$ for any $C>0, 0<\theta<1-\eta$. This set of ζ does not include the true set of "good" observations, ζ_n. Lemma A.9 states that $h^{1-\theta}(\hat{\sigma}_\zeta^2-\hat{\sigma}_{\zeta_n}^2)$ diverges to positive infinity uniformly in values of ζ in the set. Since the function $(\hat{\sigma}_\zeta^2-\hat{\sigma}_{\zeta_n}^2)$ is zero at ζ_n, the considered set of ζ values cannot contain a minimizer in the limit.

In combination, all minimizers, $\zeta\in\mathcal{M}_n$, satisfy $\max_{\zeta\in\mathcal{M}_n}\#(\zeta\cap\zeta_n^c) = \mathrm{O}_{\mathsf{P}}(h^\theta)$. □

A.4. Main Result

Proof of Theorem 3.4. (*a*) Theorem 3.3, using the Assumptions 3.1, 3.2, and 3.3, gives that $\#(\zeta\cap\zeta_n^c) = \mathrm{O}_{\mathsf{P}}(h^\theta)$ for any $0<\theta<1$. Hence, consider minimizers ζ so that $\#(\zeta\cap\zeta_n^c)\leq g_n = Ch^\theta$. Then, by Lemma A.1(*b*, *c*) using Assumptions 3.1(*iii*) and 3.3(*ia*, *id*, *iii*), we have that $\mathcal{S}_{g_n} = \mathrm{O}_{\mathsf{P}}(h^{\theta+\eta-1})L(n)$. Since $L(n)$ is slowly varying, $\eta<1/2$ and $\theta>0$ is arbitrary, we have $\mathcal{S}_{g_n} = \mathrm{o}_{\mathsf{P}}(1)$, see Karamata (1930, p. 45).

For any minimizer $\hat{\sigma}_\zeta^2-\hat{\sigma}_{\zeta_n}^2\leq 0$. Thus we need to show that $\hat{\sigma}_\zeta^2-\hat{\sigma}_{\zeta_n}^2\geq -\epsilon h^{-1/2}$ with large probability for any small $\epsilon>0$. Lemma A.8, using Assumption 3.3(*iv*) and the fact that $\mathcal{S}_{g_n} = \mathrm{o}_{\mathsf{P}}(1)$, gives the lower bound

$$h(\hat{\sigma}_\zeta^2-\hat{\sigma}_{\zeta_n}^2)/\sigma^2 \geq \{1+\mathrm{o}_{\mathsf{P}}(1)\}\sum_{i\in\zeta\cap\zeta_n^c}\varepsilon_i^2 - \{1+\mathrm{o}_{\mathsf{P}}(1)\}\sum_{i\in\zeta^c\cap\zeta_n}\varepsilon_i^2 + \mathrm{O}_{\mathsf{P}}(1), \tag{A.21}$$

where all remainder terms are uniform in ζ. First, we bound $\sum_{i\in\zeta\cap\zeta_n^c}\varepsilon_i^2\geq 0$. Second, as $\max_{i\in\zeta_n}\varepsilon_i^2 = \mathrm{O}_{\mathsf{P}}(m_n^2)$ and $m_n^2 = \mathrm{O}_{\mathsf{P}}(n^\eta)L(n)$ by Assumption 3.3(*ib*, *id*), while using $\#(\zeta^c\cap\zeta_n)\leq Ch^\theta$, we get, uniformly in ζ,

$$\sum_{i\in\zeta^c\cap\zeta_n}\varepsilon_i^2 \leq (\max_{i\in\zeta_n}\varepsilon_i^2)\{\#(\zeta^c\cap\zeta_n)\} = \mathrm{O}_{\mathsf{P}}(h^{\theta+\eta})L(n). \tag{A.22}$$

Thus, for $\eta<1/2$ and small $\theta>0$, we get $\sum_{i\in\zeta^c\cap\zeta_n}\varepsilon_i^2 = \mathrm{o}_{\mathsf{P}}(h^{1/2})$. It follows that $0\geq\hat{\sigma}_\zeta^2-\hat{\sigma}_{\zeta_n}^2\geq\mathrm{o}_{\mathsf{P}}(h^{-1/2})$.

(*b*) We show that for all ζ so that $\#(\zeta\cap\zeta_n^c)\leq Ch^\theta$, we have

$$\mathcal{D}_\zeta = |(\sum_{i\in\zeta}x_{in}x_{in}')^{1/2}(\hat{\beta}_\zeta-\beta) - (\sum_{i\in\zeta_n}x_{in}x_{in}')^{1/2}(\hat{\beta}_{\zeta_n}-\beta)| = \mathrm{o}_{\mathsf{P}}(1).$$

Lemma A.7 using Assumption 3.3(*iv*) and $\mathcal{S}_{g_n} = \mathrm{o}_{\mathsf{P}}(1)$ gives that

$$\mathcal{D}_\zeta^2 = |A_\zeta - A_{\zeta_n}|^2 \leq \Big(\sum_{i\in\zeta\cap\zeta_n^c}\varepsilon_i^2 + \sum_{i\in\zeta^c\cap\zeta_n}\varepsilon_i^2\Big)\mathrm{O}_{\mathsf{P}}(\mathcal{S}_{g_n}) + \mathrm{o}_{\mathsf{P}}(1). \tag{A.23}$$

For the first sum, we use (A.21) in part (a) to bound

$$\{1+\mathrm{o}_{\mathsf{P}}(1)\}\sum_{i\in\zeta\cap\zeta_n^c}\varepsilon_i^2 \leq h(\hat{\sigma}_\zeta^2-\hat{\sigma}_{\zeta_n}^2)/\sigma^2 + \{1+\mathrm{o}_{\mathsf{P}}(1)\}\sum_{i\in\zeta^c\cap\zeta_n}\varepsilon_i^2 + \mathrm{O}_{\mathsf{P}}(1), \tag{A.24}$$

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

where all remainder terms are uniform in ζ. Insert the bound $\hat{\sigma}^2_{\zeta} - \hat{\sigma}^2_{\zeta_n} \leq 0$ and use that $\{1 + \mathrm{o}_{\mathsf{P}}(1)\}^{-1} = 1 + \mathrm{o}_{\mathsf{P}}(1)$ and $\{1 + \mathrm{o}_{\mathsf{P}}(1)\}\{1 + \mathrm{o}_{\mathsf{P}}(1)\} = 1 + \mathrm{o}_{\mathsf{P}}(1)$ while $\{1 + \mathrm{o}_{\mathsf{P}}(1)\}\mathrm{O}_{\mathsf{P}}(1) = \mathrm{O}_{\mathsf{P}}(1)$ to get the further bound

$$\sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 \leq \{1 + \mathrm{o}_{\mathsf{P}}(1)\} \sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 + \mathrm{O}_{\mathsf{P}}(1). \tag{A.25}$$

Insert this bound into (A.23) and use $\mathcal{S}_{g_n} = \mathrm{o}_{\mathsf{P}}(1)$ to get

$$\mathcal{D}^2_{\zeta} \leq \Big(\sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 \Big) \mathrm{O}_{\mathsf{P}}(\mathcal{S}_{g_n}) + \mathrm{o}_{\mathsf{P}}(1). \tag{A.26}$$

As noted above, $\mathcal{S}_{g_n} = \mathrm{O}_{\mathsf{P}}(n^{\theta+\eta-1})L(n)$ and $\sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 = \mathrm{O}_{\mathsf{P}}(h^{\theta+\eta})L(n)$ in (A.22). In combination, $\mathcal{D}^2_{\zeta} = \mathrm{O}_{\mathsf{P}}(h^{2\theta+2\eta-1})\{L(n)\}^2 + \mathrm{o}_{\mathsf{P}}(1)$. Noting that $L(n)$ is a slowly varying function, $\eta < 1/2$ and that $\theta > 0$ can be chosen small, we get $\mathcal{D}^2_{\zeta} = \mathrm{o}_{\mathsf{P}}(1)$. □

B. ON EXTREME AND INTERMEDIATE QUANTILES

We verify Assumption 3.3(i) for some common distributions. We let $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. from a symmetric, unbounded distribution F with extremes $\varepsilon_{(n)} = \max_{1 \leq i \leq n} \varepsilon_i$ and $\varepsilon_{(1)} = \min_{1 \leq i \leq n} \varepsilon_i$. We write $a_n \sim b_n$ if $a_n/b_n \to 1$.

B.1. Extreme Quantiles

If F has exponential tails, we can establish Assumption 3.3(ib, id) with $\eta = 0$ as follows. It suffices to show that $\varepsilon_{(n)}/a_n \to 1$ in probability for some increasing sequence a_n of logarithmic rate, so that $a_n = \mathrm{O}(n^{\eta})L(n)$ for $\eta = 0$ and a slowly varying function L. In that case, $\varepsilon_{(1)}/a_n \to -1$ by symmetry so that $\max\{\varepsilon^2_{(n)}, \varepsilon^2_{(1)}\}/\min\{\varepsilon^2_{(n)}, \varepsilon^2_{(1)}\} \to 1$ and Assumption 3.3(ib, id) follows. We can check the sufficient condition using the following multiplicative strong law of large numbers.

LEMMA B.1 (Galambos, 1978, Thm. 4.4.4). *Let* $a_n = \inf\{y : \mathsf{F}(y) \geq 1 - 1/n\}$. *Then,* $\varepsilon_{(n)}/a_n \to 1$ *a.s., if and only if, for any* $k > 1$,

$$\sum_{n=3}^{\infty} \{1 - \mathsf{F}(ka_n)\} < \infty. \tag{B.1}$$

Example B.1. Let F be standard normal. Condition (B.1) is satisfied and $a_n \sim \sqrt{2\log(n)}$ (DasGupta, 2008, Ex. 8.13). Thus, $\varepsilon^2_{(n)} \sim 2\log(n)$ a.s.

Example B.2. Let F be standard Laplace. This symmetric distribution has $\mathsf{F}(x) = 1 - \exp(-x)/2$ for $x \geq 0$ so that $a_n = \mathsf{F}^{-1}(1 - 1/n) = -\log(2/n)$ for $n > 2$. Thus, $1 - \mathsf{F}(ka_n) = (2/n)^k/2$ for $n > 2$. Since $\sum_{n=3}^{\infty} n^{-k} < \infty$ for $k > 1$ then condition (B.1) is satisfied. We note that $a_n \sim \log n$, so that $\varepsilon_{(n)} \sim \log n$ a.s.

Example B.3. Let F be double geometric with $\mathsf{f}(x) = (1-p)^{|x|-1}p/2$ for $x \in \mathbb{Z}\backslash\{0\}$, so that $\mathsf{F}(x) = 1 - (1-p)^x/2$ for $x \in \mathbb{N}$ and $a_n = \lceil \log(2/n)/\log(1-p) \rceil$ for $n > 2$ where $\lceil \cdot \rceil$

is the ceiling. Note, $a_n \sim \log n$. We note that this distribution is not of an extremal type. To see this, modify Example 1.7.15 for the geometric distribution in Leadbetter et al. (1982). To apply Lemma B.1 note that $\lceil x \rceil > x$, so that $a_n > \log(2/n)/\{2\log(1-p)\} = \tilde{a}_n$ for $n > 2$. Thus, $1 - \mathsf{F}(ka_n) \leq 1 - \mathsf{F}(k\tilde{a}_n) = (2/n)^k/2$ and the argument is completed as in Example B.2.

If F has polynomial tail behavior, so that $\eta > 0$, we need a different argument.

Example B.4. Let F be t_d with $d > 4$ degrees of freedom. The extremal quotient $\varepsilon_{(n)}/\varepsilon_{(1)}$ converges to a non-degenerate, positive distribution with median 1 (Gumbel and Keeney, 1950). Assumption 3.3(*ib*) follows. Next, $1 - \mathsf{F}(x) \sim C_d x^{-d}$ for $x \to \infty$ and $\mathsf{F}^{-1}(1-1/n) \sim c_d n^{1/d}$ for $n \to \infty$ for some constants C_d, c_d depending on d (Soms, 1976). Thus, $\{1 - \mathsf{F}(tx)\}/\{1 - \mathsf{F}(t)\} \to x^{-d}$ for $t \to \infty$ so that $\varepsilon_{(n)} \sim n^{1/d}$ (Galambos, 1978, Thm. 2.1.1). Assumption 3.3(*id*) follows for $\eta \geq 2/d$. Thus, to get $\eta < 1/2$, we need $d > 4$.

Example B.5. Suppose that the "good" errors are standard normal while the "outlier" errors satisfy $\min_{j \notin \zeta_n} \varepsilon_j = \sqrt{2\log h} - 1$. We prove that in this case:

(a) $\mathsf{P}(\min_{j \notin \zeta_n} \varepsilon_j < \max_{i \in \zeta_n} \varepsilon_i) \to 1$ and (b) $\min_{j \notin \zeta_n} \varepsilon_j^2 = m_n^2\{1 + \mathsf{o}_\mathsf{P}(1)\}$.

The proof of each of these two results uses extreme value theory for standard normal random variables. In particular, for $\varepsilon_{(n)} = \max_{1 \leq i \leq n} \varepsilon_i$ and $\varepsilon_i \sim i.i.d. \mathsf{N}(0,1)$, we have

$$\mathsf{P}[a_n\{\varepsilon_{(n)} - b_n\} \leq x] \to \exp\{-\exp(-x)\}, \tag{B.2}$$

when $a_n = (2\log n)^{1/2}$ and $b_n = (2\log n)^{1/2} - \frac{1}{2}(2\log n)^{-1/2}(\log\log n + \log 4\pi)$, see Leadbetter et al. (1982, Thm. 1.5.3).

Proof of (a). We show that $\mathcal{P}_n = \mathsf{P}(\max_{i \in \zeta_n} \varepsilon_i \leq \min_{j \notin \zeta_n} \varepsilon_j) \to 0$. By construction, $\mathcal{P}_n = \mathsf{P}(\max_{i \in \zeta_n} \varepsilon_i \leq \sqrt{2\log h} - 1)$. Standardize so that

$$\mathcal{P}_n = \mathsf{P}\{a_h(\max_{i \in \zeta_n} \varepsilon_i - b_h) \leq a_h(\sqrt{2\log h} - 1 - b_h)\}.$$

Use the expression for a_n, b_n in (B.2), but evaluated at h, to get

$$\mathcal{P}_n = \mathsf{P}\{a_h(\max_{i \in \zeta_n} \varepsilon_i - b_h) \leq x_h)\} \quad \text{where} \quad x_h = -\sqrt{2\log h} + \frac{1}{2}(\log\log h + \log 4\pi).$$

Note that $x_h \to -\infty$. Therefore, $\mathcal{P}_n \to 0$ by (B.2).

Proof of (b). We prove that $\mathcal{T}_n = (\min_{j \notin \zeta_n} \varepsilon_j^2/m_n^2) - 1$ vanishes. The "outlier" errors satisfy $\min_{j \notin \zeta_n} \varepsilon_j = (\sqrt{2\log h} - 1)$ so that $\min_{j \notin \zeta_n} \varepsilon_j^2 = (\min_{j \notin \zeta_n} \varepsilon_j)^2 = (\sqrt{2\log h} - 1)^2$. Hence,

$$\mathcal{T}_n = \frac{\min_{j \notin \zeta_n} \varepsilon_j^2}{m_n^2} - 1 = \Big(\frac{2\log h}{m_n^2} - 1\Big) + \frac{1 - 2\sqrt{2\log h}}{m_n^2}.$$

From Example B.1, we have that $\max_{i \in \zeta_n} \varepsilon_i^2 \sim 2\log h$ *a.s.* By symmetry, we also get $\min_{i \in \zeta_n} \varepsilon_i^2 \sim 2\log h$. Thus, $m_n^2 \sim 2\log h$. Insert above to get that $\mathcal{T}_n$ vanishes *a.s.*

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

B.2. Intermediate Quantiles

We now consider the Assumption 3.3(*ic*, *iiib*) concerning intermediate quantiles. We consider general (right-continuous) distribution functions F and choose the lower quantile function $\mathsf{F}^{-1}(p)=\inf\{x:\mathsf{F}\ge p\}$ as the inverse. This is left-continuous, so that $\mathsf{F}\{\mathsf{F}^{-1}(p)\}\ge p$.

LEMMA B.2. *Let* $\varepsilon_1,\dots,\varepsilon_n$ *be i.i.d. with distribution function* F *with* $\inf\{x:\mathsf{F}(x)>0\}=-\infty$. *Let* $n\to\infty$ *and* $0<\rho<1$. *Define* $C_n=\mathsf{F}^{-1}(n^{\rho-1}/\log n)/\mathsf{F}^{-1}(n^{-1}\log n)$. *For any* $C_\rho<1$ *so that* $\limsup_{n\to\infty}C_n\le C_\rho$ *then* $\varepsilon_{(n^\rho)}/\varepsilon_{(1)}\le C_\rho+\mathrm{o}_{\mathsf{P}}(1)$.

Proof. Apply Theorem 1.8.1 in Leadbetter et al. (1982) with $v_n=\mathsf{F}^{-1}(n^{-1}\log n)$ so that $n\mathsf{F}(v_n)\ge\log n\to\infty$ shows that $\mathsf{P}\{\varepsilon_{(1)}>v_n\}\to\exp(-\infty)=0$. Noting that v_n is negative, we get further that $\mathsf{P}\{\varepsilon_{(1)}/v_n\ge 1\}\to 1$.

Lemma 1 in Chibisov (1964) with $a_n=\mathsf{F}^{-1}(n^{\rho-1}/\log n), x=1$, $b_n=0$, $k_n=n^\rho$ and $u_n(x)=\{n\mathsf{F}(a_nx+b_n)-k_n\}/k_n^{1/2}$ shows that $\mathsf{P}\{\varepsilon_{(k_n)}\le a_nx+b_n\}-\Phi\{u_n(x)\}\to 0$. In our case, $u_n(x)=n^{\rho/2}\{(\log n)^{-1}-1\}\to-\infty$, so that $\Phi\{u_n(x)\}\to 0$ and $\mathsf{P}\{\varepsilon_{(k_n)}\le a_n\}\to 0$. Noting that a_n is negative, we get further that $\mathsf{P}\{\varepsilon_{(k_n)}/a_n>1\}\to 0$.

Let $\epsilon>0$ be given. Consider the set $A_n=\{\varepsilon_{(k_n)}/\varepsilon_{(1)}\le C_\rho+\epsilon\}$. We must show that $\mathsf{P}(A_n)\to 1$. Rewrite $A_n=\{(\varepsilon_{(k_n)}/a_n)<(C_\rho+\epsilon)(v_n/a_n)(\varepsilon_{(1)}/v_n)\}$. Let $B_n=\{\varepsilon_{(1)}/v_n\ge 1\}$ and $D_n=\{\varepsilon_{(k_n)}/a_n<1\}$, so that $\mathsf{P}(B_n),\mathsf{P}(D_n)\to 1$, noting that v_n,a_n are negative as found above. By assumption, $\limsup_{n\to\infty}a_n/v_n\le C_\rho$. Thus, $\forall\epsilon>0$ then $a_n/v_n\le C_\rho+\epsilon$ for large n. Hence, $(C+\epsilon)v_n/a_n\ge 1$. Thus, A_n holds on $B_n\cap C_n$, so that $\mathsf{P}(A_n)\ge\mathsf{P}(B_n\cap C_n)\to 1$. □

Example B.6. Let F be standard normal. By Mill's ratio, $x\Phi(x)\sim-\varphi(x)$ for $x\to-\infty$, so that $\log(-x)\sim\log\varphi(x)-\log\Phi(x)$. Apply for $x=\Phi^{-1}(s_n^{-1})$ with $s_n\to\infty$ to get $2\log\{-\Phi^{-1}(s_n^{-1})\}\sim-\log(2\pi)-\{\Phi^{-1}(s_n^{-1})\}^2+2\log s_n$. Since $\log\{-\Phi^{-1}(s_n^{-1})\}=\mathrm{o}\{\Phi^{-1}(s_n^{-1})\}$ then the previous asymptotic equivalence implies $2\log s_n\sim\{\Phi^{-1}(s_n^{-1})\}^2$. Thus, $\Phi^{-1}(s_n^{-1})\sim-(2\log s_n)^{1/2}\to\infty$ for $s_n\to\infty$. We find, for $0<\rho<1$, that $C_n=\{\log(n^{\rho-1}/\log n)/\log(n^{-1}\log n)\}^{1/2}\sim(1-\rho)^{1/2}=C_\rho<1$. Assumption 3.3(*ic*) follows by Lemma B.2. Example B.1 shows that $\varepsilon_{(n)}^2\sim 2\log n=\mathrm{O}(n^\eta)L(n)$ a.s. with $\eta=0$. Thus, $\forall\delta>0$, $\exists\rho<1-\eta=1$, so that $\varepsilon_{(n^\rho)}^2/\varepsilon_{(1)}^2\le C_\rho+\mathrm{o}_{\mathsf{P}}(1)<\delta$ and Assumption 3.3(*iiia*) follows.

Example B.7. Let F be Laplace. Then $\mathsf{F}(x)=\exp(x)/2$ for $x<0$ and $\mathsf{F}^{-1}(\psi)=\log(2\psi)$ for $\psi<1/2$ Thus, $C_n=\log(2n^{\rho-1}/\log n)/\log(2n^{-1}\log n)$ so that $C_n\sim 1-\rho=C_\rho<1$. Assumption 3.3(*ic*, *iiia*) follow by Lemma B.2.

Example B.8. Let F be double geometric. Then $\mathsf{F}(x)=(1-p)^{-x}/2$ for $x\in-\mathbb{N}$ and $\mathsf{F}^{-1}(\psi)\sim-\log(2\psi)/\log(1-p)$ for $\psi\to 0$. Assumption 3.3(*ic*, *iiia*) follow by Lemma B.2 since $C_n\sim\{\log(2n^{\rho-1}/\log n)\}/\{\log(2n^{-1}\log n)\}\sim 1-\rho=C_\rho<1$.

Example B.9. Let F be the t_d with d degrees of freedom, so that $\mathsf{F}^{-1}(\psi)\sim-c_d\psi^{-1/d}$ for $\psi\to 0$ and some constant c_d depending on d (Soms, 1976). Thus, for any $0<\rho<1$, we get $C_n\sim\{(n^{\rho-1}/\log n)/(n^{-1}\log n)\}^{-1/d}=n^{-\rho/d}(\log n)^{2/d}\to 0$. Thus, by Lemma B.2 we have that $\varepsilon_{(n^\rho)}/\varepsilon_{(1)}$ vanishes for any ρ.

Example B.10. Let $(-X)^{-\omega}$ with $\omega>0$ be gamma distributed with shape and inverse scale of $\nu>0$. Then $\mathsf{F}(x)=\mathsf{P}(X\le x)=\mathsf{P}(Z\le z)$, where $Z=\nu(-X)^{-\omega}$ and $z=\nu(-x)^{-\omega}$.

Since Z is gamma with shape ν and unit scale, then $\mathsf{F}(x) = \gamma(\nu,z)/\Gamma(\nu)$, where γ and Γ are the lower incomplete and the complete gamma function, respectively. For large $-x$ then z is small. We get $\mathsf{F}(x) \sim z^{\nu}/\{\nu\Gamma(\nu)\}$ for small $z > 0$ (Gradshteyn and Ryzhik, 1965, 8.354.1), so that $\mathsf{F}(x) \sim (-x)^{-\nu\omega}\nu^{\nu-1}/\Gamma(\nu)$ for large $-x$. Thus, $\mathsf{F}^{-1}(\psi) \sim -\{\psi\Gamma(\nu)/\nu^{\nu-1}\}^{-1/(\nu\omega)}$ for $\psi \to 0$. Following Lemma B.2, we find for $0 < \rho < 1$ that $C_n = \{(n^{\rho-1}/\log n)/(n^{-1}\log n)\}^{-1/(\nu\omega)} = (\log n)^{2/(\nu\omega)} n^{-\rho/\nu\omega} \to 0$ for large n so that $\varepsilon_{(n^{\rho})}/\varepsilon_{(1)}$ vanishes for any ρ.

C. HETEROSCEDASTIC EXAMPLE

Let $z = x^{-\omega}$ be gamma distributed with shape and inverse scale of $\nu = p/2$ and some $\omega > 2$. Let ε given x, and therefore also given z, be $\mathsf{N}(0,1/z)$. We will require that $p > 4$ so that x, ε have the fourth moments needed for heteroscedastic inference.

We show that ε is t_p distributed. Using a gamma integral, the density is found to be

$$\begin{aligned}\mathsf{f}_{\varepsilon}(\varepsilon) &= \int_0^{\infty} \frac{1}{\sqrt{2\pi/z}} \exp(-z\varepsilon^2/2) \frac{\nu^{\nu}}{\Gamma(\nu)} z^{\nu-1} \exp(-\nu z) dz \\ &= \frac{\nu^{\nu}}{\Gamma(\nu)\sqrt{2\pi}} \int_0^{\infty} z^{\nu-1+1/2} \exp\{-z(\nu+\varepsilon^2/2)\} dz \\ &= \left\{\frac{\nu^{\nu}}{\Gamma(\nu)\sqrt{2\pi}}\right\}\left\{\frac{\Gamma(\nu+1/2)}{(\nu+\varepsilon^2/2)^{\nu+1/2}}\right\} = \frac{\Gamma\{(p+1)/2\}}{\Gamma(p/2)\sqrt{\pi p}}(1+\varepsilon^2/p)^{-(p+1)/2}.\end{aligned} \qquad \textbf{(C.1)}$$

We show that $x = z^{-1/\omega}$ has a bounded density so that Assumption 3.1(*iii*) is satisfied through Example 4.2. By the change-of-variable formula with mapping $z \mapsto z^{-1/\omega} = x$, inverse mapping $x \mapsto x^{-\omega}$ and Jacobean $\omega x^{-\omega-1}$, we get that x has density

$$\mathsf{f}_x(x) = \mathsf{f}_z(x^{-\omega})\omega x^{-\omega-1} = \frac{\omega\nu^{\nu}}{\Gamma(\nu)} x^{-\omega\nu-1}\exp(-\nu x^{-\omega}).$$

The density is positive and continuous for $x > 0$ with $\mathsf{f}(x) \to 0$ for $x \to 0$ since the exponential function dominates the power function. Thus, the density is bounded.

We show that $\mathsf{E}x^4 < \infty$ so that Assumption 3.2(*i*) is satisfied by the Law of Large Numbers and x has the required moments. With $\nu = p/2 > 2$ and $\omega > 2$ we get

$$\begin{aligned}(\mathsf{E}x^4)^{\omega/2} \le \mathsf{E}x^{2\omega} = \mathsf{E}(1/z^2) &= \frac{\nu^{\nu}}{\Gamma(\nu)} \int_0^{\infty} \frac{1}{z^2} z^{\nu-1}\exp(-\nu z) dz \\ &= \left\{\frac{\nu^{\nu}}{\Gamma(\nu)}\right\}\left\{\frac{\Gamma(\nu-2)}{\nu^{\nu-2}}\right\} = \frac{\nu^2}{(\nu-1)(\nu-2)} < \infty.\end{aligned}$$

We study the tail behavior of x required in Assumption 3.3. First, we show that $x^2_{(n)} = \mathsf{O}_{\mathsf{P}}(m_n^2)$. Consider n i.i.d. repetitions of x, ε. We have that $\max_{1\le i\le n}\varepsilon_i^2 \sim n^{2/p}$ since ε_i is t_p and using Example B.4. Thus, we show $\mathcal{P}_n = \mathsf{P}(\max_{1\le i\le n} x_i^2 \le n^{2/p}) \to 1$. Exploiting the i.i.d. structure, we get

$$\mathcal{P}_n = \mathsf{P}\cap_{1\le i\le n}(x_i^2 \le n^{2/p}) = \{\mathsf{P}(x_1^2 \le n^{2/p})\}^n = \exp\{n\log\mathsf{P}(x_1^2 \le n^{2/p})\}. \qquad \textbf{(C.2)}$$

Exploiting that $z = x^{-\omega}$ where $y = \nu z$ is gamma with shape $\nu = p/2$ and scale 1 gives

$$\mathsf{P}(x_1^2 \le n^{2/p}) = \mathsf{P}(z \ge n^{-\omega/p}) = \mathsf{P}(y \ge \nu n^{-\omega/p}) = \frac{1}{\Gamma(\nu)}\int_{\nu n^{-\omega/p}}^{\infty} y^{\nu-1}\exp(-y)dy.$$

Expand the gamma integral (Gradshteyn and Ryzhik, 1965, 8.354.2) to get

$$\mathsf{P}(x_1^2 \leq n^{-2/p}) = 1 - \frac{(\nu n^{-\omega/p})^\nu}{\nu\Gamma(\nu)} + \mathrm{o}\{(n^{-\omega/p})^\nu\} = 1 - \frac{\nu^{\nu-1} n^{-\omega/2}}{\Gamma(\nu)} + \mathrm{o}(n^{-\omega/2}).$$

We find that $\mathsf{P}(x_1^2 \leq n^{-2/p}) = 1 + \mathrm{o}(n^{-1})$ when $\omega > 2$. Insert in (C.2) and expand the logarithm as $n\log\{1 + \mathrm{o}(n^{-1})\} = \mathrm{o}(1)$ to see that $\mathcal{P}_n \to 1$ when $\omega > 2$.

Second, we show that $\forall 0 < \delta < 1, \exists 0 < r < 1 - \eta$: $x^2_{(n-\lfloor n^r \rfloor)}/x^2_{(n)} \leq \delta\{1 + \mathrm{o}_\mathsf{P}(1)\}$. Since $\max_{1\leq i\leq n} \varepsilon_i^2 \sim n^{2/p}$ we can choose $\eta = 2/p$, so that $1 - \eta > 0$ whenever $p > 2$. Example B.10 with $\nu = p/2$ shows that $x^2_{(n-\lfloor n^r \rfloor)}/x^2_{(n)}$ vanishes in probability for any $0 < r < 1$ and in particular for $0 < r < 1 - \eta$ as desired.

D. VARIATION OF THE ASSUMPTIONS

Theorem 3.2 shows that the "good" observations are consistently selected as long as the smallest "outlier" squared error, $\min_{j\notin\zeta_n} \varepsilon_j^2$, diverges. This result allows for "good" errors with both bounded or unbounded support. Theorem 3.3 improves the consistency rate of Theorem 3.2, while Theorem 3.4 provides an asymptotic expansion of the estimators. Theorems 3.3 and 3.4 require that the largest "good" squared error, $\max_{i\in\zeta_n} \varepsilon_i^2$, diverges. Hence, these two results apply for "good" errors with unbounded support. Here, we investigate how far we can get with bounded "good" errors.

Assumption D.1. Suppose

(i) **"Good" errors**: $1/(\max_{i\in\zeta_n} \varepsilon_i^2) = \mathrm{O}_\mathsf{P}(1)$;
(ii) **"Outlier" errors**: $(\max_{i\in\zeta_n} \varepsilon_i^2)/(\min_{j\notin\zeta_n} \varepsilon_j^2) = \mathrm{o}_\mathsf{P}(1)$;
(iii) **Regressors**: Let $|x_{in}|$ have order statistics $x_{(1)} \leq \cdots \leq x_{(n)}$ satisfying $x^2_{(n)} = \mathrm{o}_\mathsf{P}(\min_{j\notin\zeta_n} \varepsilon_j^2)$;
(iv) **Infeasible OLS estimator**: $(\hat{\beta}_{\zeta_n} - \beta)'(\sum_{i\in\zeta_n} x_{in}x'_{in})(\hat{\beta}_{\zeta_n} - \beta) = \mathrm{O}_\mathsf{P}(1)$.

Note that in Assumption D.1, we have that part (*iv*) restates Assumption 3.3(*iv*).
We start with a variation of Lemma A.1.

LEMMA D.1 (Variation of Lemma A.1). *Suppose Assumption D.1(iii). Then,* $\forall C > 0$: $\mathcal{R}_{g_n} = \mathrm{o}_\mathsf{P}(h)$ *for* $g_n \leq Ch/\min_{j\notin\zeta_n} \varepsilon_j^2$.

Proof of Lemma D.1. Bound $\mathcal{R}_{g_n} \leq g_n x^2_{(n)}$. Since $g_n \leq Ch/\min_{j\notin\zeta_n} \varepsilon_j^2$, by construction and $x^2_{(n)} = \mathrm{o}_\mathsf{P}(\min_{j\notin\zeta_n} \varepsilon_j^2)$ by Assumption D.1(*iii*), then $\mathcal{R}_{g_n} = \mathrm{o}_\mathsf{P}(h)$. □

We take the following results as stated in Appendix A.3:
Lemma A.1(*c*) stands using Assumption 3.1(*iii*).
Lemmas A.2, A.3, A.4, A.5, A.6, stand using no Assumptions.
Lemmas A.7, A.8 stand using Assumption D.1(*iv*) that replaces Assumption 3.3(*iv*).

LEMMA D.2 (Variation of Lemma A.9). *Suppose Assumptions 3.1(iii), D.1. Then,* $\min_{\zeta:1\leq\#(\zeta\cap\zeta_n^c)\leq hC/\min_{j\notin\zeta_n}\varepsilon_j^2} h(\hat{\sigma}^2_\zeta - \hat{\sigma}^2_{\zeta_n}) \to \infty$ *in probability.*

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

Proof of Lemma D.2. Let # be shorthand for $\#(\zeta^c \cap \zeta_n) = \#(\zeta \cap \zeta_n^c)$.

We consider $1 \leq \# \leq g_n$ where $g_n = hC/\min_{j \notin \zeta_n} \varepsilon_j^2$. We get that $\mathcal{S}_{g_n} = \mathcal{R}_{g_n} \mathrm{O}_{\mathsf{P}}(n^{-1})$ by Lemma A.1(*c*) using Assumption 3.1(*iii*). Further $\mathcal{R}_{g_n} = \mathrm{o}_{\mathsf{P}}(h)$ by Lemma D.1 using Assumption D.1(*iii*). Thus, $\mathcal{S}_{g_n} = \mathrm{o}_{\mathsf{P}}(1)$. Then, Lemma A.8 using Assumption D.1(*iv*) shows

$$h(\hat\sigma_\zeta^2 - \hat\sigma_{\zeta_n}^2)/\sigma^2 \geq \{1 + \mathrm{o}_{\mathsf{P}}(1)\} \sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 - \{1 + \mathrm{o}_{\mathsf{P}}(1)\} \sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 + \mathrm{O}_{\mathsf{P}}(1), \tag{D.1}$$

where all remainder terms are uniform in ζ. We show that the lower bound diverges. Insert the following bounds. For $i \in \zeta \cap \zeta_n^c$, then $\varepsilon_i^2 \geq \min_{j \notin \zeta_n} \varepsilon_j^2$. For $i \in \zeta^c \cap \zeta_n$, then $\varepsilon_i^2 \leq \max_{i \in \zeta_n} \varepsilon_i^2$. Further, the sums in (D.1) have the same number of elements $\# = \#(\zeta^c \cap \zeta_n) = \#(\zeta \cap \zeta_n^c)$. Thus,

$$h(\hat\sigma_\zeta^2 - \hat\sigma_{\zeta_n}^2)/\sigma^2 \geq \big(\min_{j \notin \zeta_n} \varepsilon_j^2\big)\{1 + \mathrm{o}_{\mathsf{P}}(1)\}\# - \big(\max_{i \in \zeta_n} \varepsilon_i^2\big)\{1 + \mathrm{o}_{\mathsf{P}}(1)\}\# + \mathrm{O}_{\mathsf{P}}(1).$$

Take common factor $(\min_{j \notin \zeta_n} \varepsilon_j^2)\#$ to get

$$h(\hat\sigma_\zeta^2 - \hat\sigma_{\zeta_n}^2)/\sigma^2 \geq \Big[\{1 + \mathrm{o}_{\mathsf{P}}(1)\} - \Big(\frac{\max_{i \in \zeta_n} \varepsilon_i^2}{\min_{j \notin \zeta_n} \varepsilon_j^2}\Big)\{1 + \mathrm{o}_{\mathsf{P}}(1)\}\Big]\big(\min_{j \notin \zeta_n} \varepsilon_j^2\big)\# + \mathrm{O}_{\mathsf{P}}(1).$$

Since $\# \geq 1$ by construction in this Lemma and $(\max_{i \in \zeta_n} \varepsilon_i^2)/(\min_{j \notin \zeta_n} \varepsilon_j^2)$ vanishes while $\min_{j \notin \zeta_n} \varepsilon_j^2$ diverges due to Assumption D.1(*i, ii*), then $h(\hat\sigma_\zeta^2 - \hat\sigma_{\zeta_n}^2)/\sigma^2$ diverges with large probability. □

THEOREM D.3 (Variation of Theorem 3.3). *Suppose Assumptions 3.1, 3.2, D.1. Let $\mathcal{M}_n$ denote the set of minimizers ζ of $\hat\sigma_\zeta^2$. Then,* $\mathsf{P}\{\max_{\zeta \in \mathcal{M}_n} \#(\zeta \cap \zeta_n^c) = 0\} \to 1$.

Proof of Theorem D.3. First, Theorem 3.2, using Assumptions 3.1, 3.2, shows that $\max_{\zeta \in \mathcal{M}_n} \#(\zeta \cap \zeta_n^c) = \mathrm{O}_{\mathsf{P}}(h/\max_{j \notin \zeta_n} \varepsilon_j^2)$.

Second, Lemma D.2, using Assumptions 3.1(*iii*), D.1, considers estimators $\hat\sigma_\zeta^2$ for index sets ζ that contain a positive number of "outliers", in the range, $1 \leq \#(\zeta \cap \zeta_n^c) \leq Ch/\max_{j \notin \zeta_n} \varepsilon_j^2$. Such sets ζ do not include the true set of "good" observations, ζ_n. Lemma D.2, states that $h(\hat\sigma_\zeta^2 - \hat\sigma_{\zeta_n}^2)$ diverges to positive infinity uniformly in considered values of ζ. Since the function $(\hat\sigma_\zeta^2 - \hat\sigma_{\zeta_n}^2)$ is zero at ζ_n, the considered set of ζ values cannot contain a minimizer in the limit.

In combination, minimizers $\zeta \in \mathcal{M}_n$ satisfy $\mathsf{P}\{\max_{\zeta \in \mathcal{M}_n} \#(\zeta \cap \zeta_n^c) = 0\} \to 1$. □

THEOREM D.4 (Variation of Theorem 3.4). *Suppose Assumptions 3.1, 3.2, D.1. Then*

(a) $\mathsf{P}(\hat\sigma^2 = \hat\sigma_{\zeta_n}^2) \to 1$.

(b) $\mathsf{P}\{(\sum_{i \in \hat\zeta} x_{in} x_{in}')^{1/2}(\hat\beta - \beta) - (\sum_{i \in \zeta_n} x_{in} x_{in}')^{1/2}(\hat\beta_{\zeta_n} - \beta)\} \to 1$.

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

Proof of Theorem D.4.

(a) Theorem D.3, using the Assumptions 3.1, 3.2, D.1, shows that $\mathsf{P}\{\max_{\zeta\in\mathcal{M}_n}\#(\zeta\cap\zeta_n^c)=0\}\to 1$. Whenever $\max_{\zeta\in\mathcal{M}_n}\#(\zeta\cap\zeta_n^c)=0$, then the LTS estimator $\hat{\sigma}^2$ equals the OLS estimator on ζ_n. Thus, $\mathsf{P}(\hat{\sigma}^2=\hat{\sigma}^2_{\zeta_n})\to 1$.

(b) Further, whenever $\max_{\zeta\in\mathcal{M}_n}\#(\zeta\cap\zeta_n^c)=0$, then the LTS estimator $\hat{\beta}$ equals the OLS estimator on ζ_n, while $\hat{\zeta}$ equals ζ_n. Thus, the desired result follows. □

REFERENCES

Agullo, J., Croux, C., & Van Aelst, S. (2008). The multivariate least-trimmed squares estimator. *Journal of Multivariate Analysis*, 99, 311–338.

Alfons, A., Croux, C., & Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Annals of Applied Statistics*, 7, 226–248.

Atkinson, A. C., Riani, M., & Cerioli, A. (2010). The forward search: Theory and data analysis (with discussion). *Journal of the Korean Statistical Society*, 39, 117–163.

Bednarski, T., & Clarke, B. R. (1993). Trimmed likelihood estimation of location and scale of the normal distribution. *Australian Journal of Statistics*, 35, 141–153.

Berenguer-Rico, V., Johansen, S., & Nielsen, B. (2023). A model where the least trimmed squares estimator is maximum likelihood. *Journal of the Royal Statistical Society. Series B*, 85, 886–912.

Berenguer-Rico, V. , & Nielsen, B. (2023). Normality testing after outlier removal. *Econometrics and Statistics*.

Butler, R. (1982). Nonparametric interval and point prediction using data trimmed by a Grubbs-type outlier rule. *Annals of Statistics*, 10, 197–204.

Chen, X. R., & Wu, Y. H. (1988). Strong consistency of M-estimators in linear models. *Journal of Multivariate Analysis*, 27, 116–130.

Chibisov, D. M. (1964). On limit distributions for order statistics. *Theory of Probability and Its Applications*, 9, 142–147.

Čížek, P. (2005). Least trimmed squares in nonlinear regression under dependence. *Journal of Statistical Planning and Inference*, 136, 3967–3988.

Clarke, B. R. (2018). *Robustness theory and application*. John Wiley & Sons.

Croux, C. , & Rousseeuw, P. J. (1992). A class of high-breakdown scale estimators based on subranges. *Communications in Statistics. Theory and Methods*, 21, 1935–1951.

DasGupta, A. (2008). *Asymptotic theory of statistics and probability*. Springer.

Davies, L. (1990). The asymptotics of S-estimators in the linear regression model. *Annals of Statistics*, 18, 1651–1675.

Galambos, J. (1978). *The asymptotic theory of extreme order statistics*. John Wiley & Sons.

Gallegos, M. T. , & Ritter, G. (2009). Trimmed ML estimation of contaminated mixtures. *Sankhya A*, 71, 164–220.

Gradshteyn, I. S., & Ryzhik, I. M. (1965). *Table of integrals, series and products* (4th ed.). Academic Press.

Gumbel, E. J., & Keeney, R. D. (1950). The extremal quotient. *Annals of Mathematical Statistics*, 21, 523–538.

He, X., Jurečková, J., Koenker, R., & Portnoy, S. (1990). Tail behavior of regression estimators and their breakdown points. *Econometrica*, 58, 1195–1214.

Hössjer, O. (1994). Rank-based estimates in the linear model with high breakdown point. *Journal of the American Statistical Association*, 89, 149–158.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, 73–101.

Johansen, S. (1995). *Likelihood based inference on cointegration in the vector autoregressive model*. Oxford University Press.

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press

Johansen, S., & Nielsen, B. (2009). Saturation by indicators in regression models. In J. L. Castle & N. Shephard (Eds.), *The methodology and practice of econometrics: Festschrift in honour of David F* (pp. 1–36). Oxford University Press.

Johansen, S., & Nielsen, B. (2016). Asymptotic theory of outlier detection algorithms for linear time series regression models (with discussion). *Scandinavian Journal of Statistics*, 43, 321–381.

Johansen, S. , & Nielsen, B. (2019). Boundedness of M-estimators for multiple linear regression in time series. *Econometric Theory*, 35, 653–683.

Karamata, J. (1930). Sur une mode de croissance régulière des fonctions. *Mathematica (Cluj)*, 4, 38–53.

Leadbetter, M. R., Lindgren, G., & Rootzén, H. (1982). *Extremes and related properties of random sequences and processes*. Springer.

Raymaekers, J. , & Rousseeuw, P. J. (2024). The cellwise minimum covariance determinant estimator. *Journal of the American Statistical Association*, 119, 2610–2621.

Rousseeuw, P. J. (1984). Least median of squares regressions. *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, & W. Wertz (Eds.), *Mathematical statistics and applications* (pp. 283–297). Reidel.

Rousseeuw, P. J. (1994). Unconventional features of positive-breakdown estimators. *Statistics & Probability Letters*, 19, 417–431.

Rousseeuw, P. J. , & Hubert, M. (1997). Recent developments in PROGRESS. In Y. Dodge (Ed.), *L_1-statistical procedures and related topics*, vol 31 of Lecture Notes–Monograph Series (pp. 201–214). Institute of Mathematical Statistics.

Rousseeuw, P. J. , & Leroy, A. M. (1987). *Robust regression and outlier detection*. John Wiley & Sons.

Rousseeuw, P. J., Perrotta, D., Riani, M., & Hubert, M. (2019). Robust monitoring of time series with application to fraud detection. *Econometrics and Statistics*, 9, 108–121.

Rousseeuw, P. J., & van Driessen, K. (2000). An algorithm for positive-breakdown regression based on concentration steps In W. Gaul, O. Opitz, & M. Schader (Eds.), *Data analysis: Scientific modeling and practical application* (pp. 335–346). Springer Verlag.

Scholz, F. W. (1980). Towards a unified definition of maximum likelihood. *Canadian Journal of Statistics*, 8, 193–203.

Soms, A. P. (1976). An asymptotic expansion for the tail area of the t-distribution. *Journal of the American Statistical Association*, 71, 728–730.

Vandev, D. L., & Neykov, N. M. (1993). Robust maximum likelihood in the Gaussian case. In S. Morgenthaler, E. Ronchetti, & W. A. Stahel (Eds.), *New directions in data analysis and robustness* (pp. 259–264). Birkhäuser.

Víšek, J. Á. (2006). The least trimmed squares; part III: Asymptotic normality. *Kybernetika*, 42, 203–224.

Watts, V., Rootzén, H., & Leadbetter, M. R. (1982). On limiting distributions of intermediate order statistics from stationary sequences. *Annals of Probability*, 10, 653–662.

Wilms, I., & Croux, C. (2016). Robust sparse canonical correlation analysis. *BMC Systems Biology*, 10, Article 72, 13 pp.

https://doi.org/10.1017/S0266466624000343 Published online by Cambridge University Press